SEMANTIC IMAGE SEGMENTATION USING THE ICM ALGORITHM

Lazhar Khelifi, Max Mignotte

Image Processing Laboratory, DIRO, University of Montreal, Montreal, Canada
Email: khelifil@iro.umontreal.ca, mignotte@iro.umontreal.ca

ABSTRACT

Semantic image segmentation has recently become the focus of considerable interest. This task consists in assigning a predefined class label to each pixel (or pre-segmented region) in an image. To address the complexity challenge of this task, we develop, in this work, a novel and simple energy-minimization model. The proposed cost function of this model combines efficiently different global non-parametric semantic likelihood energy terms computed from the (pre-)segmented regions of the (query) image and their structural properties (location, texture, color, context and shape). To optimize our energy-based model, we use a local optimization procedure derived from the iterative conditional modes (ICM) algorithm. Experimental results on the challenging Microsoft Research Cambridge dataset (MSRC-21) clearly shows the feasibility and the merits of the proposed approach.

Index Terms— Semantic Segmentation, energy minimization model, Microsoft Research Cambridge (MRC) dataset.

1. INTRODUCTION

In the last few years there has been a growing interest in the semantic segmentation (also called scene parsing). The aim of this task is to divide image into semantic regions, such as *mountain, sky, building, tree, etc.* One challenge of scene parsing is that it combines three traditional problems; detection, segmentation and multi-label recognition in a single process [1].

As an active research area, various methods for scene parsing have been proposed in the literature. As mentioned in [2], these methods can be generally classified into three groups based on the relationships (dependencies) which are encoded between different pixels in the image. The first type contains methods which solve the pixel-labeling problem by classifying each pixel independently [3]. However, the high computational cost of these approaches and their inefficiency makes them unattractive to applications. The second type of methods is based on the pairwise Markov Random Field (MRF) or Conditional Random Field (CRF) models [4], where nodes in the graph represent the semantic label associated with a pixel, and potentials are created to define the energy of the system. Thus, a relationship between pairs of neighbouring pixels is incorporated in the graph, which encourage adjacent pixels that are similar in appearance to take the same semantic label. However, in this type of framework, the learning and inference of complex pairwise terms are often expensive. In addition, this approach is still too local and not descriptive enough to capture long-range relationships observed between adjacent regions. In the third group, pixels are grouped into segments (or super-pixels) and a single label is assigned to each group [5]. Following this approach, a probabilistic model characterizing spatial context for region annotation has been proposed in [6]. Also, we can mention the supervised image annotation method in [7] based on regional features, which considers other regions of the image as meaningful context of the current region.
The work presented in this paper aims at overcoming the drawbacks of previous techniques by proposing a simple energy-minimization model called the multi-criteria semantic segmentation model (MC-SSM). The proposed model combines efficiently different global likelihood terms either based on the spatial organization and distribution of the region semantic labels within the image or on region-based properties (location, texture, color, context and shape), and their training adequacy, in a multi-criteria cost function. To optimize our energy-model, we use a simple local optimization procedure derived from the iterative conditional modes (ICM) algorithm.

2. OUR MC-SSM MODEL

Our scene parsing procedure is performed through two steps. In the first step, a set of segments (regions) is generated by a pre-segmentation [1] algorithm called GCEBFM [11, 12]. In the second step, based on an available labeled segmentation corpus, a single class label is assigned to each region by optimizing a global fitness function that measures the *quality* of the generated solution. To this end, relevant features are extracted from these individual segments. The used features allow to capture different aspects of color (COL), texture (TEX), shape (SHA), image location (LOC), semantic contextual information (CTX) for an image region and their adequacy for a given semantic label.

Mathematically, let us assume that we have an input image I and its region segmentation $R_I = \{r_I^1, r_I^2, \ldots, r_I^m\}$ (generated by the GCEBFM algorithm) to be semantically labeled, where m represents the number of regions (r) in R_I. Let also $\mathcal{C} = \{\mathcal{I}_k, \mathcal{S}_k\}_{k \leq K}$ represents respectively a set (or a training corpus) of K images $\mathcal{I}_k$ and their corresponding semantic segmentations $\mathcal{S}_k$. In our framework, if $\mathcal{S}_\Omega$ represents the set of all possible semantically labeled segmentation maps of I (based on its partition into regions R_I) then, our semantic

[1] Another alternative to generate the segment candidates is to use over-segmentation algorithms such as; the SLIC algorithm [8], the BASS algorithm [9] or the mean shift algorithm [10] (by varying the spatial and range bandwidth parameters).

labeling problem $\hat{S}_{\mathrm{MC}} = \{s_I^1, s_I^2, \ldots, s_I^m\}$ is formulated as the result of the following multi-criteria optimization problem:

$$\hat{S}_{\mathrm{MC}} = \arg\min_{S \in \mathcal{S}_\Omega} \overline{\mathrm{MC}}\left(I, R_I, S, \{\mathcal{I}_k, \mathcal{S}_k\}_{k \leq K}\right) \quad (1)$$

with:
$$\begin{aligned}\overline{\mathrm{MC}}\left(I, R_I, S, \{\mathcal{I}_k, \mathcal{S}_k\}_{k \leq K}\right) = \\ \alpha \sum_{i=1}^{m} \mathrm{COL}\left(I, r_I^i, s_I^i, \{\mathcal{I}_k, \mathcal{S}_k\}^{s_I^i}\right) \\ + \beta \sum_{i=1}^{m} \mathrm{TEX}\left(I, r_I^i, s_I^i, \{\mathcal{I}_k, \mathcal{S}_k\}^{s_I^i}\right) \\ + \gamma \sum_{i=1}^{m} \mathrm{SHA}\left(r_I^i, s_I^i, \{\mathcal{I}_k, \mathcal{S}_k\}^{s_I^i}\right) \\ + \delta \sum_{i=1}^{m} \mathrm{LOC}\left(r_I^i, s_I^i, \{\mathcal{I}_k, \mathcal{S}_k\}^{s_I^i}\right) \\ + \lambda \sum_{i=1}^{m} \frac{1}{h} \left\{ \sum \mathrm{CTX}\left(r_I^m, s_I^i, \{\mathcal{I}_k, \mathcal{S}_k\}^{s_I^i}\right) \right\}\end{aligned}$$

where the parameters α, β, γ, δ and λ are used to weight the different terms of this energy function. COL, TEX, SHA, LOC and CTX designate respectively the different energy terms , or non-parametric distance measures, of this cost function, reflecting the adequacy of a specific semantic label (existing in the training corpus $\{\mathcal{I}_k, \mathcal{S}_k\}_{k \leq K}$) for each region of the image, in terms of its color, texture, shape, image location and semantic contextual information.

More precisely, if $\{\mathcal{C}\}^{s_I^i} = \{\mathcal{I}_k, \mathcal{S}_k\}^{s_I^i}$ denotes the set of image and associated semantic segmentation solution (belonging to the training corpus) that contains a region semantically labeled s_I^i and h represents the total number of those semantic segmentations in the corpus $\{\mathcal{C}\}^{s_I^i}$ (see Table 1).

- COL(.) is the minimum Ruzicka distance[2] between the p-bin ($p = 5^3$) normalized color histogram of r_I^i and the color histogram of each region corresponding to the semantic label assigned to r_I^i (*i.e.*, s_I^i) and existing in $\{\mathcal{C}\}^{s_I^i}$.
- TEX(.) uses the same distance previously described but based on the q-bin normalized histogram of oriented gradients (HOG) with 4 different directions and 10 amplitude values.
- LOC(.) is the minimum absolute distance, normalized in term of percentage of image height, between the height of the topmost pixel existing in the region r_I^i and the topmost pixel of each region corresponding to the semantic label assigned to r_I^i (*i.e.*, s_I^i) and existing in $\{\mathcal{C}\}^{s_I^i}$.
- SHA(.) uses the same distance previously described but based on the normalized area of the considered region.
- CTX(.) exploits the semantic contextual information around each region. More precisely, CTX(.) is the Ruzicka distance[2] between the 21-bin normalized histogram of semantic labels of r_I^i (excluding its own semantic label s_I^i) and the histogram of semantic labels of each region, existing in $\{\mathcal{C}\}^{s_I^i}$, and corresponding to the semantic label assigned to r_I^i (*i.e.*, s_I^i).

[2] $\mathrm{distance}_{\mathrm{Ruzicka}} = 1 - \sum_i [\min(P_i, Q_i) / \max(P_i, Q_i)]$

A synoptic illustration of our segmentation method is shown in Fig. 1.

3. OPTIMIZATION PROCEDURE

Our semantic segmentation model of multiple label fields is formulated as a global optimization problem incorporating a nonlinear objective function. To enable us to achieve the minimum of this energy function [see (2)], approximation approaches based on different optimization algorithms such as the exploration/selection/estimation (ESE) [13], the genetic algorithm or the simulated annealing can be exploited. These algorithms are guaranteed to find the optimal solution, but with the drawback of a huge computational time. To overcome this problem, in this work we adopt the iterated conditional modes (ICM) method proposed by Besag [14] (i.e.; a Gauss-Seidel relaxation), where pixels (semantic label of each region in our case) are updated one at a time. In our case, this algorithm turned out to be both easy to implement, fast and efficient in terms of convergence properties (the algorithm is fast converging after 10 iterations according to our experiments). The entire pseudo-code of our MC-SSM based on ICM is shown in Algorithm 1.

Algorithm 1 MC-Semantic Segmentation Model algorithm

Mathematical notation:

$\overline{\mathrm{MC}}$	Multi-criteria function
$\{V_k\}_{k \leq K}$	Set of K images
$\{S_k\}_{k \leq K}$	Set of K semantic segmentations (related to $\{V_k\}_{k \leq K}$)
$\mathcal{E}$	Set of class labels in $\{S_k\}_{k \leq K}$
$T_{\max}$	Maximal number of iterations (=100)
$\hat{S}_{best}$	Semantic segmentation result
I	Image to be labeled
R_I	Region segmentation of image I

Input: I, $\{V_k\}_{k \leq K}$, $\{S_k\}_{k \leq K}$

Output: $\hat{S}_{best}$

A. Initialization:

Segment image I into different coherent regions R_I (with the GCEBFM algorithm [11])

Assign class label for each r_i region $\in R_I$ using random element from $\mathcal{E}$

B. Steepest Local Energy Descent:

while $p < T_{\max}$ **do**
  for each r_i region $\in R_I$ **do**
    Draw a new class label y according to the uniform distribution in the set $\mathcal{E}$
    Let $R_I^{[p],\mathbf{new}}$ the new semantic segmentation map including r_i with the class label y
    Compute $\overline{\mathrm{MC}}(R_I^{[p],\mathbf{new}}, \{S_k\}_{k \leq K})$ *[see (2)]*
    if $\overline{\mathrm{MC}}(R_I^{[p],\mathbf{new}}, \{S_k\}_{k \leq K}) < \overline{\mathrm{MC}}(R_I^{[p]}, \{S_k\}_{k \leq K})$ **then**
      $\overline{\mathrm{MC}} = \overline{\mathrm{MC}}^{\mathbf{new}}$
      $R_I^{[p]} = R_I^{[p],\mathbf{new}}$
      $\hat{S}_{best} = R_I^{[p]}$
    end if
  end for
  $p \leftarrow p + 1$
end while

Table 1. Summary of the combined criteria used in our model.

Criterion	Name	Dimension
-Color-	Color histogram	125
-Texture-	Oriented gradient histogram	40
-Shape-	Pixel area	1
-Location-	Top height	1
-Context-	Context histogram	21

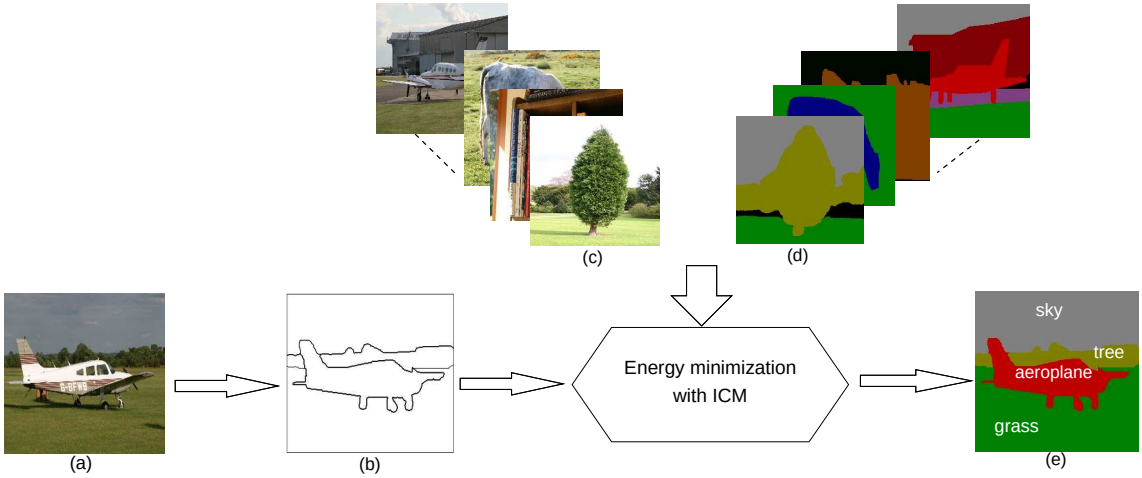

Fig. 1. Proposed system overview. (a) input image. (b) image segmentation (achieved by the GCEBFM algorithm [12]). (c) and (d) corpus of images and its semantic segmentation. (e) scene parsing result.

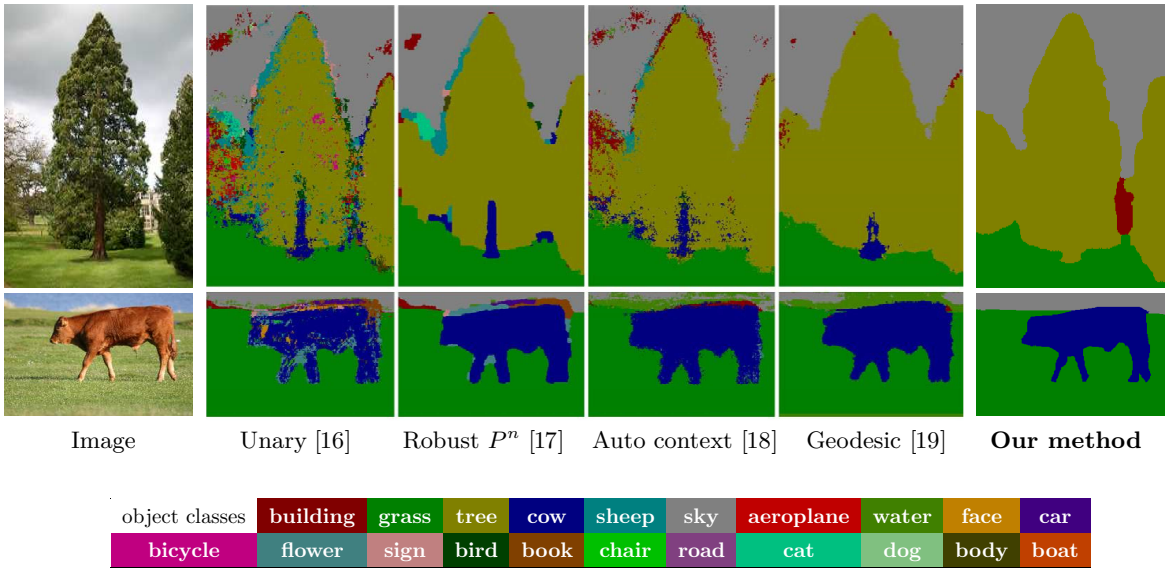

Fig. 2. Example of segmentation results obtained by our algorithm MC-SSM on two images from the MSRC-21 compared to other algorithms.

4. EXPERIMENTS

We evaluate our method on the Microsoft Research Cambridge MSRC-21 dataset [15]. The MSRC-21 dataset contains 591 color images with corresponding ground truth labelling for 21 classes (building, grass, tree, cow...). We adopt the leave-one-out evaluation strategy. Thus, for each image, we use it as a query image and we classify its region based on the rest of the images in the dataset. The image annotation performance is measured by the global accuracy,

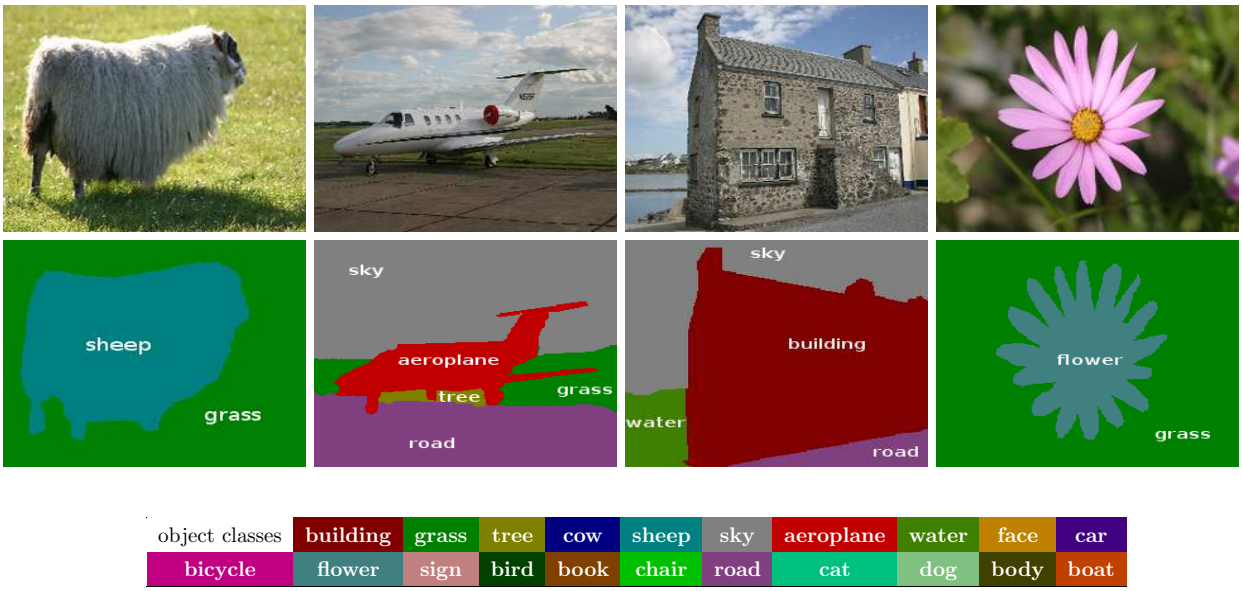

Fig. 3. Example results obtained by our MC-SSM model on the MSRC-21-class dataset (for more clarity, we have superimposed textual labels on the resulting segmentations).

which is widely used for evaluating the performances of related tasks. To guarantee the integrity of the benchmark results, the five weight parameters of our algorithm [i.e., α, β, γ, δ and λ, see (2)] are optimized on the ensemble of training images by using a local linear search procedure in the feasible ranges of parameter values ($[1:2]$) with a fixed step-size $= 10^{-2}$. We have found that $\alpha = 1.83$, $\beta = 1.53$, $\gamma = 1.55$, $\delta = 1.70$ and $\lambda = 1$, are reliable hyper-parameters for the model yielding the best accuracy value. Table 2 shows that the result achieved by our technique is comparable to the state-of-the-art methods, although is much simpler than different algorithms. Additionally, we present a qualitative comparison with other methods; Unary [16], Robust P^n [17], Auto context [18] and Geodesic [19] (see Fig. 2). Also, Fig. 3 shows other semantic segmentation results on the MSRC-21 generated by our algorithm, the whole results of the dataset are accessible on-line via this link: http://www-etud.iro.umontreal.ca/~khelifil/ResearchMaterial/mc-ssm.html. As we can notice, our multi-criteria semantic segmentation model (MC-SSM) is both simple and efficient and can be regarded as a robust alternative to complex, computationally demanding semantic segmentation models existing in the literature. Finally, it is worth mentioning that improvements can be made efficiently in our algorithm by adding other interesting invariant features (to the multi-criteria function) such as the SIFT (scale-invariant feature transform) or the LSD (line segment detector) descriptors or similarity measure between segmentations.

Table 2. Global accuracy on the MSRC-21 dataset (the higher is better).

Algorithms	**Accuracy (%)**
-SuperParsing- [20] in [21]	61.50
-SIM- [22]	69.70
-SVM-BoW - [23]	62.70
-MC-SSM-	66.07

5. CONCLUSION

In this paper, we studied the problem of semantic segmentation (called also scene parsing). Towards this goal, we proposed a novel and simple energy-minimization called the multi-criteria semantic segmentation model (MC-SSM). Our approach achieved state-of-the-art performance in the popular MSRC-21 dataset. Furthermore, we plan to extend our model by using other optimization algorithm, and to improve further the classification accuracy by incorporating others criteria.

6. REFERENCES

[1] C. Farabet, C. Couprie, L. Najman and Y. LeCun, "Learning Hierarchical Features for Scene Labeling," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 35, no. 8, pp. 1915–1929, 2013.

[2] J. Shotton and P. Kohli, "Semantic Image Segmentation," *in: K. Ikeuchi, (eds.) Computer Vision: A Reference Guide*, pp. 713–716, Springer US, 2014.

[3] 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," *Int. J. Comput. Vis.*, vol. 81, no. 1, pp. 2–23, 2009.

[4] Z. Liu, X. Li, P. Luo, C. C. Loy and X. Tang, "Semantic Image Segmentation via Deep Parsing Network," *in IEEE Int. Conf. Comp. Vis. (ICCV)*, 2015, pp. 1377–1385.

[5] N. W. Campbell, W. Mackeown, B. T. Thomas and T. Troscianko, "Interpreting image databases by region classification," *Pattern Recognit.*, vol. 30, no. 4, pp. 555–563, 1997.

[6] Z. Wang, D. D. Feng, Z. Chi and T. Xia, "Annotating Image Regions Using Spatial Context," *in Proc. of the 8 th IEEE Int. Sym. Mult.*, 2006, pp. 55–61.

[7] F. Shi, J. Wang and Z. Wang, " Region-based supervised annotation for semantic image retrieval," *Int. J. elect. Communicat.*, vol. 65, no. 11, pp. 929-936, 2011.

[8] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua and S. Susstrunk, 'SLIC super-pixels compared to state-of-the-art superpixel methods," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2274–2282, 2012.

[9] A. Rubio, L. Yu, E. Simo-Serra and F. Moreno-Noguer (in press), " A simple high performance approach to semantic segmentation," *in Int. Conf. Pattern Recognit. (ICPR)*, 2016.

[10] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 5, pp. 603–619, 2002.

[11] L. Khelifi and M. Mignotte (in press), "A novel fusion approach based on the global consistency criterion to fusing multiple segmentations," *IEEE Trans. Syst., Man, Cybern., Syst.*, Mar. 2016.

[12] L. Khelifi and M. Mignotte, "GCE-based model for the fusion of multiples color image segmentations," *23rd IEEE Int. Conf. Image Process. (ICIP)*, 2016, pp. 2574–2578.

[13] F. Destrempes, M. Mignotte, and J.-F. Angers, "A stochastic method for Bayesian estimation of hidden Markov models with application to a color model," *IEEE Trans. Image Process.*, vol. 14, no. 8, pp. 1096–1108, 2005.

[14] J. Besag, "On the statistical analysis of dirty pictures," *J. Royal Stat. Soc.*, vol. 48, no. 3, pp. 259–302, 1986.

[15] J. Shotton, J. Winn, C. Rother and A. Criminisi, "TextonBoost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation," *in: A. Leonardis, H. Bischof, A. Pinz,(eds.) ECCV 2006.* LNCS, vol. 3951, pp. 1–15. Springer, Heidelberg, 2006.

[16] L. Ladicky, C. Russell, P. Kohli and P.H.S. Torr, "Associative hierarchical random fields," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 6, pp. 1056–1077, 2013.

[17] P. Kohli, L. Ladicky and P.H.S. Torr, "Robust higher order potentials for enforcing label consistency," *Int. J. Comput. Vis.*, vol. 82, no. 3, pp. 302–324, 2009.

[18] Z. Tu and X. Bai, "Auto-context and its application to high-level vision tasks and 3D brain image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 10, pp. 1744–1757 , 2010.

[19] V. Haltakov, C. Unger and S. Ilic, "Geodesic pixel neighborhoods for 2D and 3D scene understanding," *Comput. Vis. Image Underst.*, vol. 148, pp. 164–180 , 2016.

[20] J. Tighe and S. Lazebnik, "SuperParsing: Scalable Nonparametric Image Parsing with Superpixels," *in Computer Vision âĂŞ ECCV 2010, K. Daniilidis, P. Maragos, andN. Paragios (eds.)*, Springer Berlin Heidelberg, pp. 352–365, 2010.

[21] A. Bassiouny and M. El-Saban, "Semantic segmentation as image representation for scene recognition," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, 2014, pp. 981–985.

[22] F. Briggs, X. Z. Fern, and R. Raich, "Rank-loss support instance machines for MIML instance annotation," *Proceedings of the 18th ACM SIGKDD*, 2012, pp. 534–542.

[23] F. Liu, G. Lin, and C. Shen, "CRF learning with CNN features for image segmentation," *Pattern Recognit.*, vol. 48, no. 10, pp. 2983–2992, 2015.