MULTI-LABEL ENERGY MINIMIZATION FOR OBJECT CLASS SEGMENTATION

Camille Couprie

New York University,
Dept. of Computer Science, Courant Institute,
755 Broadway, New York, USA

ABSTRACT

The task of associating a semantic class to the objects present in an image is challenging because this problem involves the joint segmentation and recognition of the objects. In this work, we use a recent approach embedding several optimization algorithms into a common framework named Power watershed to perform this task. We show how the fast watershed algorithm can be used to minimize an energy function for which the minimizer corresponds to the desired object class segmentation. Higher order potentials are then added to improve label consistency. We also demonstrate that the random walker algorithm can be successfully applied to semantic class segmentation problems. Comparisons with the Graph Cuts algorithm show that the proposed approaches yield better segmentation results, obtained up to twelve times faster on a very challenging indoor scenes dataset.

***Index Terms*—** Image processing, Object class segmentation, Graph-based optimization, Graph cuts, Random walker, Watershed.

1. INTRODUCTION

The purpose of object class segmentation is to label each pixel of a scene with the category of the object of which it belongs. A popular approach for this problem is the use of Markov Random Fields or Conditional Random Fields [1, 2, 3, 4, 5]. To define appropriate weights on the graphical models then created, we have nowadays the opportunity to use additional information than simply the images themselves. For example, the work of [6] proposes to use depth information extracted from the kinect device, to refine object class segmentation using Graph Cuts [7, 8]. The expression of the object class segmentation problem in an energy formulation is very useful to the use of additional extensions exploiting different information.

For example, [9] showed how to minimize energies with high order cliques using graph cuts, leading to a better enforcement of label consistency inside objects for object recognition [10], and the addition of co-occurrence statistics [11].

However, there are several drawbacks when using graph cuts, such as long computation times, block artifacts, and the fact that Graph Cuts only lead to local minima when the number of labels is larger than two.

In this paper, we propose alternative optimization methods as efficient energy minimizers for object class segmentation. Recently, Couprie *et al.* introduced the Power watershed method [12], which is related to the Graph cuts [7], Watershed [13], and Random walker [14] methods for image segmentation. Although this technique was introduced in the context of image segmentation, the authors described how the method could be used as an optimization method for various functionals, such as image filtering [15], and surface reconstruction [16]. In the present work we show that the Power watershed method is well-suited to address the object class segmentation problem, and present a successful way to deal with higher order energy terms.

2. REPRESENTATION OF ENERGY FUNCTIONS WITH GRAPHS

Since the Power watershed is defined on a graph, we begin by casting the semantic segmentation problem formulation in discrete terms. A graph consists of a pair $G = (V, E)$ with vertices $v \in V$ and edges $e \in E \subseteq V \times V$ with cardinalities $n = |V|$ and $m = |E|$. An edge, e, spanning two vertices, v_i and v_j, is denoted by e_{ij}. The goal of this work being to label all the vertices of G, given some vertices of known labels, we split V in two disjoints sets of vertices, noted V_k and V_u, and containing the vertices of known and unknown labels respectively. A weighted graph assigns a (typically non-negative and real) value to each edge called a weight. The weight of an edge e_{ij} is denoted by w_{ij}. We denote by $|S|$ the cardinality of a set S.

2.1. Power Watersheds

The generalized Power Watershed energy is given by

$$\arg\min_x \sum_{\substack{e_{ij}\in E,\\ (v_i,v_j)\in V_u\times V_u}} {w_{ij}}^p |x_i - x_j|^q + \sum_{\substack{e_{ij}\in E,\\ v_i\in V_u, v_j\in V_k}} {w_{ij}}^p |x_i - y_j|^q \tag{1}$$

Acknowledgement: the author would like to thank Nathan Silberman for sharing his software and his helpful comments.

where y represents a measured configuration and x represents the target configuration. In the first term of this equation, w_{ij} can be interpreted as a weight on the gradient of the target configuration, such that the first term penalizes any unwanted high-frequency content in x and essentially forces x to vary smoothly within an object, while allowing large changes across the object boundaries. The second term enforces fidelity of x to a specified configuration y, w_{ij} being weights enforcing that fidelity (See Fig. 1(a)).

The different values of the real numbers p and q lead to different algorithms for optimizing the energy. When the power of the weight, p, is finite, and the exponent $q = 1$, Eq. (1) leads to a binary solution x, that can be deduced using network flow techniques, also known as Graph cuts [8]. When $q = 2$, the unique solution to Eq. (1) may be obtained by solving a linear system of equation, the corresponding algorithm for image segmentation being known as the Random walker [14]. As described in [12], when the exponent p tends toward infinity, the cut obtained when minimizing the energy is a watershed cut [17], which has been shown to be equivalent to Maximum Spanning Forests [17] (MSF). Furthermore, [12] presents an algorithm – called Power watershed – to compute the unique watershed that optimizes the energy for $q = 2$ and $p \to \infty$.

2.2. Multi-class segmentation using Power watersheds

For the problem of multi-region segmentation, where the number of different regions L is larger than two, the energy expressed in (1) has to be written under a different form. We suppose that we have a set of known labels noted $y_1, ..., y_{|V_k|}$ taking their values between 1 and L. The problem is to find a labeling s defined as the argument maximum of L pairwise labellings $x^{(1)}, x^{(2)}, ..., x^{(L)}$ given by

$$\arg\min_{x=[x^{(1)},x^{(2)},...,x^{(L)}]} \sum_{l=1}^{L} \left[\sum_{\substack{e_{ij}\in E,\\ (v_i,v_j)\in V_u\times V_u}} {w_{ij}}^p |x_i^{(l)} - x_j^{(l)}|^q \right. \\ \left. + \sum_{\substack{e_{ij}\in E,\\ v_i\in V_u, v_j\in V_k}} {w_{ij}}^p |x_i^{(l)} - y_j^{(l)}|^q \right] \tag{2}$$

where $\forall v_i \in V_k, \forall l \in 1, ..., L$, $y_i^{(l)} = \begin{cases} 1 & \text{if } y_i = l, \\ 0 & \text{otherwise.} \end{cases}$ The final labeling s is given, for every $v_i \in V_u$, by

$$s_i = \arg\max\left(x_i^{(1)}, ..., x_i^{(L)}\right). \tag{3}$$

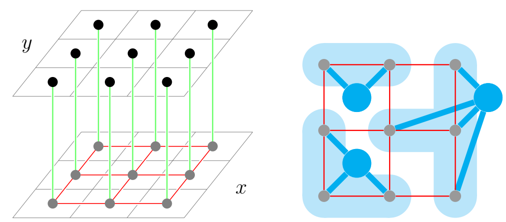
Property 1. *If $q = 2$, the optimal solution x^* of (2) satisfies $\forall v_i \in V_u$, $x_i^{(1)} + x_i^{(2)} + ... + x_i^{(L)} = 1$.*

As explained in [14], the solution to the combinatorial Dirichlet problem $\min_x \sum_{e_{ij}\in E} w_{ij}(x_i - x_j)^2$, subject to boundary conditions – some values of x enforced to be 0 or 1 – corresponds to the probability of a random walker reaching vertices marked to 1 before vertices marked 0. As probabilities, they sum to one. In the rest of this paper, the Random walker algorithm solves (2) with $q = 2$ and $p = 1$, and the Power watershed algorithm solves (2) with with $q = 2$ and $p \to \infty$.

2.3. Multi-class segmentation using higher order potentials

In the particular case where all weights are different, and $p \to \infty$, the labeling x produced by the Power watershed algorithm is binary [18]. The algorithm corresponds in this case to a simple maximum spanning forest computation [17]. We explore this particular case in this section in order to optimize a more general function.

Following the work of [10], we can enforce global consistency by introducing higher-order potentials to the energy function. Each clique c corresponds in practice to a set of nodes $\{v_{c_1}, v_{c_2}, ..., v_{c_{|c|}}\}$ of a super-pixel extracted from an over-segmentation of the image. The set of cliques is denoted in what follows by $\mathcal{S}$.



(a) Graph with pairwise and unary weights

(b) Additional Higher Order Cliques

Fig. 1. Graph built. (a): The set of grey nodes represent V_u, to which labels x are going to be associated. The set of black nodes represents V_k, for which known values y are given. The red edges represent the pairwise edges (first term of (2)), the green edges represent unary edges (second term of (2)). (b): An example of graph construction for higher order potential enforcement. Three groupings of nodes are given. All nodes inside these groupings are connected to an additional clique node (in blue) by additional blue edges.

Consequently, the labeling s of (3) may be obtained directly by solving

$$\arg\min_s \sum_{\substack{e_{ij}\in E,\\ v_i\in V_u, v_j\in V_k}} \psi(s_i, y_j) + \sum_{\substack{e_{ij}\in E,\\ (v_i,v_j)\in V_u\times V_u}} \varphi(s_i, s_j) + \sum_{c\in\mathcal{S}} \phi(s_c), \tag{4}$$

where

$$
\begin{aligned}
\varphi(s_i, s_j) &= \mathbf{1}(s_i \neq s_j) w_{ij}^p, \\
\psi(s_i, y_j) &= \mathbf{1}(s_i \neq y_j) w_{ij}^p, \\
\phi(s_c) &= \mathbf{1}(\exists (v_i, v_j) \in c \times c \text{ such that } s_i \neq s_j) w_c^p.
\end{aligned} \tag{5}
$$

We need to define additional nodes and edges to the original graph G in order to solve the new problem (4). Let V_h be a set of additional nodes, each node v_c of V_h is associated with one clique c. Let E_h be a set of $|V_u|$ additional edges, each edge e_{ic} of E_h links a node of V_u to a node of V_h and has a weight initialized to w_{ic} (See Figure 1 for an illustration). The algorithm for solving (4) is given in Algorithm 1.

Algorithm 1: Maximum Spanning Forest algorithm for the optimization of the energy (4) with higher order potentials and $p \to \infty$

Data: A weighted graph $G(V, E)$, where $V = V_u \cup V_k \cup V_h$, and $E = E_u \cup E_k \cup E_h$. Nodes of V_u and V_h have unknown potentials initially.
Result: A labeling s associating a label to each vertex.
Sort the edges of E by decreasing order of weight.
while *any node has an unknown potential* **do**
- Find the edge e_{ij} in E of maximal weight;
- **if** v_i *or* v_j *have unknown label values* **then**
  - Merge v_i and v_j into a single node, such that when the value for this merged node becomes known, all merged nodes are assigned the same value of s and considered known.
- **if** v_i *and* v_j *have known different label values and* $e_{ij} \in E_h$ **then**
  - Set all weights of the corresponding clique to 0.

3. APPLICATION TO OBJECT CLASS SEGMENTATION

We used for our experiments the NYU depth dataset of Silberman and Fergus [6], composed of 2347 triplets of images, depth images, and ground truth labeled images. The objects cover twelve categories. Most datasets used for object class segmentation present the objects centered into the images, under nice lightening conditions. The NYU depth dataset aims to develop joint segmentation and classification solutions to an environment that we are likely to encounter in the everyday life. This indoor dataset contains scenes of offices, stores, rooms of houses containing many occluded objects unevenly lightened. In this work, we are using the predictions of a classifier used in [6] and trained using the features described in [19]. It is worth mentioning that this classifier has a very good accuracy on the scene category database (81%) and only poor results on the NYU dataset (53%), demonstrating the very challenging nature of this dataset.

The NYU dataset is provided with ten possible splits of the train and test images. We use the first split in our tests. We realize in our experiments a comparison between the current state-of-the-art method for performing object-class segmentation on the NYU depth dataset, that is to say the Graph cuts method, and three algorithms: the Random Walker – to our knowledge, applied here for the first time to semantic scene segmentation–, Power watersheds and our Maximum Spanning Forest algorithm using higher order potentials.

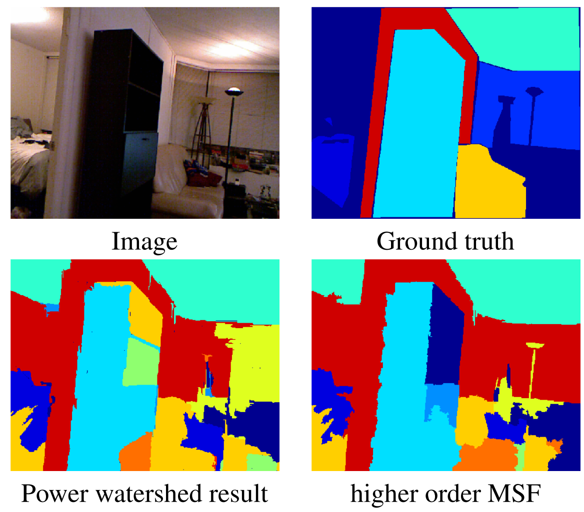


Fig. 2. Comparison of results obtained with Power watersheds and higher order MSF. The color legend is given in Fig.3.

For the four methods, the pairwise weights correspond to a metric inversely function of the image gradient that also takes the depth image into account. The unary terms correspond to learned appearance model from images and depth maps. All the details about the parameters used for the unary and pairwise weights are given in [6] (we chose the parameters giving the best results). The choice of higher order cliques was motivated to enforce a local consistency in regions generated by the hierarchical segmentation method of [20]. The weights w_c were set to the size of each segment, $w_c = |c|$, where $|c|$ was normalized.

Results are reported in Table 1. We quantify the *segmentation* accuracy in the results using three different standard segmentation measures used in [21] , namely Rand Index (RI), Global Consistency Error (GCE), and Variation of Information (VoI). Good segmentation results are associated with high RI, low GCE and low VoI. The *classification* accuracy is the recognition accuracy computed from the confusion matrix. We give two measures of classification accuracy: the accuracy per class, given by the mean diagonal of the confusion matrix, and the accuracy per pixel, computed as the ratio of correctly classified pixels versus the total number of pixels of the dataset.

The results in Table 1 demonstrate the superior segmen-

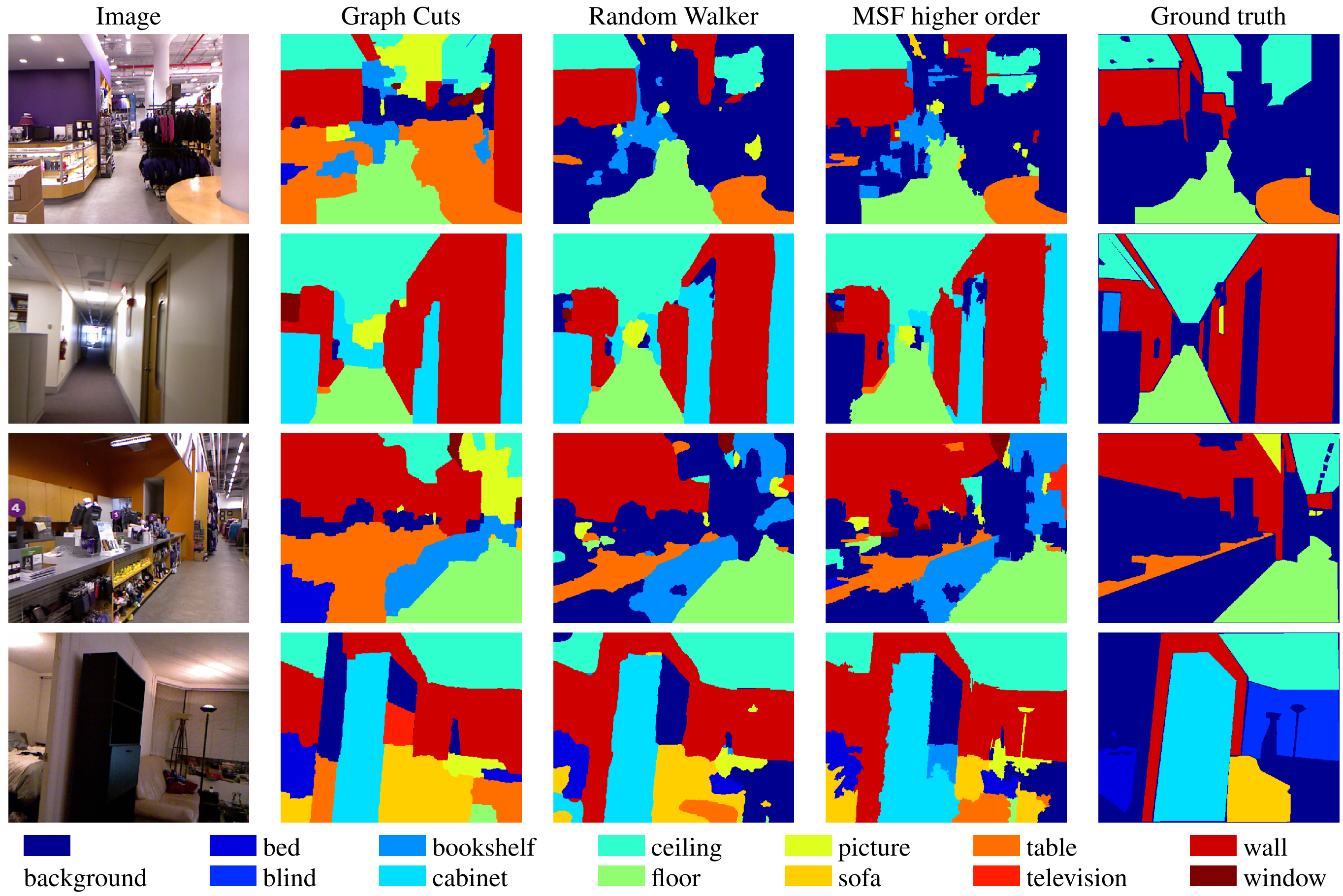


Fig. 3. Results obtained with Graph cuts, Random walker and MSF with higher order potentials on the NYU depth dataset.

tation performance of the Power watershed – as well as the Random walker and MSF employed with higher order terms – over the graph cuts method. In terms of classification accuracy, Graph Cuts are more accurate than other methods for recognizing rare objects. However, Random Walker, followed by MSF and Power Watershed, tend to produce better delineation of objects, manage to recover the correct class in the majority of the cases. Visually, a lack of regularization in the Power watershed results leads to small label inconsistency artifacts, that are overcome by the higher order MSF method (See Fig.3). Examples of results are presented in Fig. 3.

In terms of computation times, the Power watershed method is six times faster than the graph cut method. Both methods were implemented on CPU in C, using Matlab interfaces. We used for our tests a standard PC with a processor Intel Core i3-2100 CPU at 3.10GHz. Segmenting a 640×480 image using Graph Cuts takes in average more than 5 seconds, and less than one second using Power watershed. The MSF with higher order terms is the fastest algorithm, because no pass for locating area of same weights is necessary. One could argue that it requires an additional pass for computing a super pixels segmentation, but this step – that takes less than one second – is necessary to all methods as detailed in [6].

4. CONCLUSION

In this work we introduced a novel core of methods for object class segmentation, bringing several breakthroughs to this problem: we showed how a greedy procedure can optimize exactly a meaningful multi-label energy defined in a graph, that model the problem in an appropriate way. In particular, the higher order potential employed here allow us to enforce labels consistency accurately. The speed of this watershed based procedure is more than ten times faster than the classical graph cut method used in this context. Although Graph Cuts are more accurate to recognize objects belonging to classes that are under-represented in the dataset used in this work, the results obtained with the proposed approaches reach a better classification and segmentation accuracy. Future work will aim to improve the current system by enforcing co-occurrence statistics and building hierarchical schemes [22, 23].

		Graph Cuts	Rand. Walk.	Pow. Wat.	High. ord. MSF
Segmentation	Mean RI	57.3	66.0	**66.2**	66.0
	Median RI	59.4	67.0	**67.3**	66.5
	Std. dev. RI	9.1	13.8	13.2	13.8
	Mean GCE	0.48	**0.35**	0.36	0.36
	Mean VoI	3.1	**2.4**	2.5	2.5
Classif.	Per class	**56.4**	45.4	47.8	46.5
	Per pixels	46.2	**56.2**	53.1	53.7
	Time (s)	5.11	4.7	0.85	**0.4**

Table 1. Segmentation, classification accuracy, and computation times. Mean accuracy computed between the segmentation/classification masks and the ground truth images from the NYU depth database. See the text for more details.

5. REFERENCES

[1] Joseph Tighe and Svetlana Lazebnik, "Superparsing: Scalable nonparametric image parsing with superpixels," in *11th European Conference on Computer Vision, Heraklion, Greece, September 5-11, 2010, Proceedings, ECCV (5)*, 2010, pp. 352–365.

[2] F. Ning, D. Delhomme, Yann LeCun, F. Piano, Léon Bottou, and Paolo Emilio Barbano, "Toward automatic phenotyping of developing embryos from videos," *IEEE Transactions on Image Processing*, vol. 14, no. 9, pp. 1360–1371, 2005.

[3] Stephen Gould, Richard Fulton, and Daphne Koller, "Decomposing a scene into geometric and semantically consistent regions," in *IEEE 12th International Conference on Computer Vision, ICCV 2009, Kyoto, Japan, September 27 - October 4, 2009*, 2009, pp. 1–8.

[4] Xuming He and Richard S. Zemel, "Learning hybrid models for image annotation with partially labeled data," in *Advances in Neural Information Processing Systems 21, Vancouver, British Columbia, Canada, December 8-11*, 2008, pp. 625–632.

[5] Philipp Krahenbuhl and Vladlen Koltun, "Efficient inference in fully connected crfs with gaussian edge potentials," in *Neural Information Processing Systems*, 2011.

[6] Nathan Silberman and Rob Fergus, "Indoor scene segmentation using a structured light sensor," in *3DRR Workshop, ICCV'11*, 2011.

[7] Yuri Boykov and Marie-Pierre Jolly, "Interactive graph cuts for optimal boundary and region segmentation of objects in n-d images," in *ICCV*, 2001, pp. 105–112.

[8] Yuri Boykov, Olga Veksler, and Ramin Zabih, "Fast approximate energy minimization via graph cuts," *IEEE Trans. on Pattern Anal. Mach. Intell.*, vol. 23, no. 11, pp. 1222–1239, 2001.

[9] Pushmeet Kohli, M. Pawan Kumar, and Philip Torr, "P3 & beyond: Solving energies with higher order cliques," in *Conference in Computer Vision and Pattern Recognition*, 2007.

[10] Pushmeet Kohli, Lubor Ladicky, and Philip Torr, "Robust higher order potentials for enforcing label consistency," in *Conference in Computer Vision and Pattern Recognition*, 2008.

[11] Lubor Ladicky, Chris Russell, Pushmeet Kohli, and Philip H. S. Torr, "Graph cut based inference with co-occurrence statistics," in *Proceedings of the 11th European conference on Computer vision: ECCV (5)*, Berlin, Heidelberg, 2010, pp. 239–253, Springer-Verlag.

[12] Camille Couprie, Leo J. Grady, Laurent Najman, and Hugues Talbot, "Power watershed: A unifying graph-based optimization framework," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 7, pp. 1384–1399, 2011.

[13] L. Vincent and P. Soille, "Watersheds in digital spaces: An efficient algorithm based on immersion simulations," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 13, pp. 583–598, 1991.

[14] Leo Grady, "Random walks for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 11, pp. 1768–1783, 2006.

[15] Camille Couprie, Leo J. Grady, Laurent Najman, and Hugues Talbot, "Anisotropic diffusion using power watersheds," in *International Conference on Image Processing, ICIP'10, Sept. 26-29, Hong Kong, China*, 2010, pp. 4153–4156.

[16] Camille Couprie, Xavier Bresson, Laurent Najman, Hugues Talbot, and Leo J. Grady, "Surface reconstruction using power watershed," in *10th International Symposium on Mathematical Morphology, ISMM'11, Verbania-Intra, Italy, July 6-8*, 2011, pp. 381–392.

[17] Jean Cousty, Gilles Bertrand, Laurent Najman, and Michel Couprie, "Watershed cuts: Minimum spanning forests and the drop of water principle," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 8, pp. 1362–1374, 2009.

[18] Camille Couprie, Leo J. Grady, Laurent Najman, and Hugues Talbot, "Power watersheds: A new image segmentation framework extending graph cuts, random walker and optimal spanning forest," in *IEEE International Conference on Computer Vision, ICCV'09, Kyoto, Japan, Sept. 27 - Oct. 4*, 2009, pp. 731–738.

[19] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'06), 17-22 June, New York, USA*, 2006, pp. 2169–2178.

[20] Silvio Jamil F. Guimarães, Jean Cousty, Yukiko Kenmochi, and Laurent Najman, "An efficient hierarchical graph based image segmentation," 2012.

[21] Allen Yang, John Wright, Yi Ma, and Shankar Sastry, "Unsupervised segmentation of natural images via lossy data compression," *Computer Vision and Image Understanding*, vol. 110, no. 2, pp. 212–225, May 2008.

[22] Victor Lempitsky, Andrea Vedaldi, and Andrew Zisserman, "A pylon model for semantic segmentation," in *Advances in Neural Information Processing Systems*, 2011.

[23] Clément Farabet, Camille Couprie, Laurent Najman, and Yann LeCun, "Scene parsing with multiscale feature learning, purity trees, and optimal covers," in *Proc. of International Conference on Machine Learning*, 2012.