MC-SSM: Nonparametric Semantic Image Segmentation With the ICM Algorithm

Lazhar Khelifi and Max Mignotte

***Abstract*—In the last few years, there has been considerable interest in scene parsing. This task consists of assigning a predefined class label to each pixel (or pre-segmented region) in an image. To best address the complexity challenge of this task, first, we propose a new geometric retrieval strategy to select nearest neighbors from a database containing fully segmented and annotated images. Then, we introduce a novel and simple energy-minimization model. The proposed cost function of this model combines efficiently different global nonparametric semantic likelihood energy terms. These terms are computed from the (pre-)segmented regions of the (query) image and their structural properties (location, texture, color, context, and shape). Different from the traditional approaches, we use a simple and local optimization procedure derived from the iterative conditional modes algorithm to optimize our energy-based model. Experimental results on two challenging datasets: 1) microsoft research Cambridge dataset and 2) Stanford background dataset demonstrate the feasibility and the success of the proposed approach. Compared to existing annotation methods that require training classifiers for each object and learning many parameters, our method is easy to implement, has a few parameters, and combines different criteria.**

***Index Terms*—Image processing, semantic image segmentation, energy minimization model, iterative conditional modes (ICM), global consistency error (GCE).**

I. Introduction

SCENE parsing, also called semantic image segmentation, has been attracting considerable interest in the last few years. This task aims to divide an image into semantic regions or objects [1], such as *mountain, sky, building, etc*. The main challenge of scene parsing is that it combines three traditional problems; detection [2], segmentation [3] [4] [5] and multi-label recognition [6] in a single process [7]. This task aims to assign an object class label from a predetermined label set to each pixel (or super-pixel)[1] in an input image [8], [9].

As an active research area, various methods for scene parsing have been proposed in the literature. The existing methods fall into three categories. The first one is the parametric approach that uses machine learning techniques to learn compact parametric models for categories of interest in the image. Following this strategy, we can learn parametric classifiers to recognize objects (for example, building or sky) [12]. In this field several deep learning techniques [13]–[15] have been applied to semantic segmentation, for example a parametric scene parsing algorithm based on the convolutional neural networks (CNNs) [16] has been presented recently in [7]. In this algorithm, CNNs aim to learn strong features and classifiers to discriminate the local visual subtleties. The second is the nonparametric approach which aims to label the input image by matching parts of images to similar parts in a large dataset of labeled images. Here, the category classifier learning is replaced in general by a Markov random field in which unary potentials are computed by nearest-neighbor retrieval [12]. In the third category, a nonparametric model is integrated with a parametric model [17]. In this context, a quasi-parametric (hybrid) method, which integrates K-nearest neighbor (KNN)-based nonparametric method and CNN-based parametric method, has been proposed in [18]. Inspired by this method, a new automatic nonparametric image parsing framework towards leveraging the advantages of both parametric and nonparametric methodologies, has been also developed in [19].

Although the parametric approach has achieved great success on the scene parsing, all current parametric methods have certain limitations in terms of training time [20]. Another source of the problem is the retraining of models as new training dataset is added. This updating task is necessary and even important for such task, by the fact that the number of object labels in such parsing models is limited. However, the number of objects is actually unlimited in the real world. In contrast, for nonparametric approaches, no special accommodation is required when the vocabulary of semantic category labels is expanded, because there is no need to retrain category models when we add a new data [12].

To cope with these aforementioned problems related to parametric methods, in this paper, following the nonparametric approach, we propose a simple energy-minimization model called the multi-criteria semantic segmentation model (MC-SSM). The

Manuscript received April 24, 2018; revised September 30, 2018 and November 27, 2018; accepted December 20, 2018. Date of publication January 9, 2019; date of current version July 19, 2019. This work was supported in part by the National Science and Engineering Research Council of Canada (NSERC), and in part by the Tunisia's Universitary Mission in North-America (MUTAN). The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Jian Zhang. *(Corresponding author: Lazhar Khelifi.)*

The authors are with the Department of Computer Science and Operations Research, Faculty of Arts and Sciences, University of Montreal, Montreal, QC H3C 3J7, Canada (e-mail: khelifil@iro.umontreal.ca; mignotte@iro.umontreal.ca).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2019.2891418

[1] In general, super-pixel is defined as a set of connected pixels having similar appearance [10], [11].

potential aim of this new model is to take advantages of the complementarity of different criteria or features. Thus, 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. In order to optimize our energy model, we use a simple local optimization procedure derived from the iterative conditional modes (ICM) algorithm.

The contributions of this work can be itemized as follows:

- This work presents a new geometric retrieval strategy to select nearest neighbors from a database containing fully segmented and annotated images. This strategy is based on a new criterion called global consistency error (GCE). An important benefit of this criterion is the ability to find matches between the region map or the segmentation of the input image and the region map of each image in the dataset. This is particularly well suited and useful for finding a relatively smaller and interesting set of images instead of using the entire training set.
- Inspired by the recent progress in the field of image segmentation, we propose a novel energy-minimization based approach called the multi-criteria semantic segmentation model (MC-SSM). This new approach aims to assign to each region a single class label based on a global fitness function, while limiting the number of parameters. Furthermore, by combining different types of features into the energy or the objective function, our model integrates more information about the object possibly present in the scene.
- Generally, semantic segmentation models require large datasets to train high-accuracy classifiers. On the contrary, the proposed model is dedicated to small datasets which characterized by a limited number of available images.
- We report evaluations of our method on two challenging datasets; Microsoft research Cambridge dataset (MSRC-21) and Stanford background dataset (SBD), which are publicly available. The obtained results demonstrate the feasibility and the success of the proposed approach.

In the following, the paper is structured as follows: A literature review concerning the nonparametric approach for scene parsing is presented in Section II. Then our semantic segmentation model is discussed in detail in Section III. Experimental results and comparisons with existing scene parsing methods are illustrated in Section IV. In this section, our method is validated on two publicly available databases. A summary of our method and discussion of the conclusions are presented in Section V.

II. Related Work

In nonparametric scene parsing approach, 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 [21] [22]. Following this strategy, we can mention the system proposed by Liu *et al.* [23], which selects a subset of the nearest neighbors for an input image, using a large dataset that contains fully annotated images. In this system, a dense correspondence is established between the query image and each of the nearest neighbors using the SIFT flow algorithm [24]. Then, the annotations are transferred from the retrieved subset to the input image using a Markov random field (MRF) defined over pixels. However, the high computational cost of these types of methods and their inefficiency makes them unattractive to applications. The second type of methods is based on the pairwise MRF or conditional random field (CRF) models [25], 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 neighboring pixels is incorporated in the graph, which encourages 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 [26]. Following this approach, an efficient nonparametric image parsing method called Superparsing [27] has been proposed by Tighe *et al.*, in this method, an MRF is applied over super-pixels instead of pixels, then labels are transferred from a set of neighbor images to the input image based on super-pixels similarity. Also, Zand *et al.* [28] have proposed recently an ontology-based semantic image segmentation using mixture models and multiple CRFs. By doing so, the problem of image segmentation is then reduced to that of a classification task where CRFs individually classify image regions into appropriate labels for each visual feature. Moreover, Xie *et al.* [8] have proposed a new semantic image segmentation method addressing multiscale features and contextual information. In their work, an over-segmentation is applied on a given image to generate various small-scale segments, and a segment-based classifier with a CRF model are used to generate large-scale regions, then the features of regions are exploited to train a region-based classifier.

It is important to note that, there are two main questions that need to be asked when we follow the nonparametric image parsing approach, which are: a) How to retrieve some similar images from a training dataset for a query test image; b) How to parse the test image with the retrieved images by transferring the annotation associated with the retrieved images to the query image [29]. In this work, to solve the first problem, we propose a new selection process based on a new criterion called global consistency error. For the second issue, as shown in the preliminary work [30], we propose a novel energy-minimizing framework, which aims to assign to each region a single class label based on a global fitness function.

III. Model Description

As mentioned in Section I, our main aim is to decompose an image I into an unknown number (K) of geometric regions, and then to identify their categories (i.e., tree, building, mountain, etc.) by iteratively optimizing a multi-criteria energy function

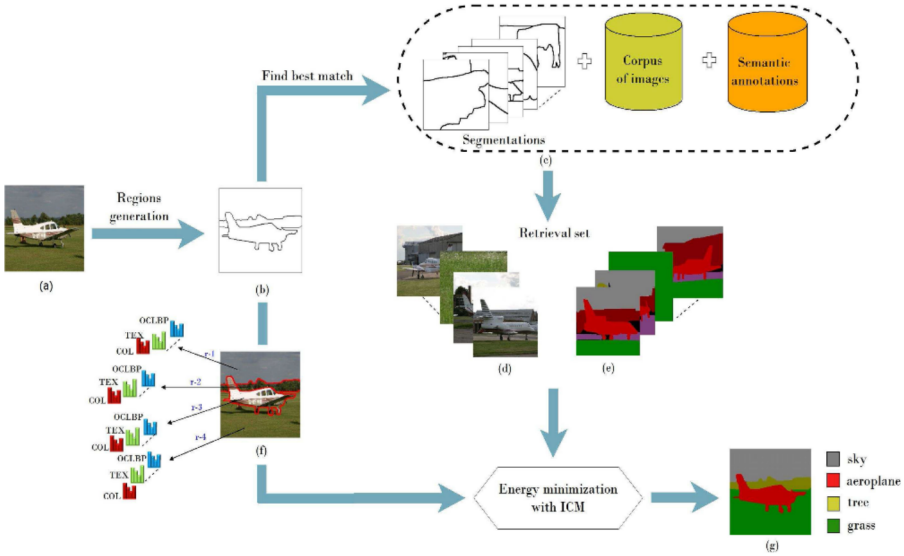

Fig. 1. System overview. Given an input image (a), we generate its set of regions with the GCEBFM algorithm (b), we retrieve similar images from the full dataset (c) using the GCE criterion, we extract different features both for the input image (f) and the retrieved images (d). Based on the labeled segmentation corpus (e), a single class label is assigned to each region (g) using energy minimization based on the ICM.

that evaluates the quality of the solution at hand. Fig. 1 illustrates the proposed system overview, which consists of following four steps: i) Region generation creates a set of regions (i.e., objects) for a given input image. ii) Geometric retrieval set selects a subset of images from the entire dataset, by a new matching scheme based on the global consistency error (GCE) measure. iii) Region features extract different types of features for each region, including color, texture, shape, image location and semantic contextual information (both for the input image and the retrieval set). iv) Image labeling assigns each region with an object class label by using an energy minimization scheme. In the following subsections, each step of our model is discussed in detail.

A. Regions Generation

In this first step, a set of segments (regions) is generated by a new pre-segmentation algorithm called GCEBFM [31], [32]. This novel algorithm aims to obtain a final refined segmentation by combining multiple and eventually weak segmentation maps generated by the standard K-means algorithm. This algorithm is applied on 12 different color spaces in order to ensure variability in the segmentation ensemble, those are, YCbCr, TSL, YIQ, XYZ, h123, P1P2, HSL, LAB, RGB, HSV, i123, and LUV. This new algorithm has been adopted in our work mainly for two reasons; Firstly, as it has been mentioned in [31], this fusion algorithm remains simple to implement, perfectible, by incrementing the number of segmentations to be fused, and general enough to be applied to different types of images. Secondly, previously published studies [10] that use predefined super-pixels,[1] generated by an over-segmentation, provide boundaries which are often inconsistent with the true region boundaries, and in most cases, objects are segmented into many regions, making an accurate decomposition of the image impossible. On the contrary, this algorithm aims to generate large regions which allow us to derive global properties for each region (see Section III-C), and on the other hand, to reduce the complexity and the memory requirement of the full model. Also, it is important to note that the performance of this new fusion model was evaluated on the Berkeley dataset [33] including various segmentations given by humans (in [31] more explanations are given about this new algorithm). Fig. 2 shows examples of initial segmentation ensemble and fusion results of an input image chosen from the MSRC-21 dataset [34].

B. Geometric Retrieval Set

In our method, we follow the hypothesis, indicating that using a subset of images which are similar to the query image, instead of using the entire dataset, is more useful for the labeling task. Note that it could be meaningful to labeling an object as a tree if we search for the nearest neighbors in images of gardens and eliminate views from indoor scenes. With the aim of finding a relatively smaller and interesting set of images instead of using the entire training set, we use a new criterion called global consistency error (GCE) to find matches between

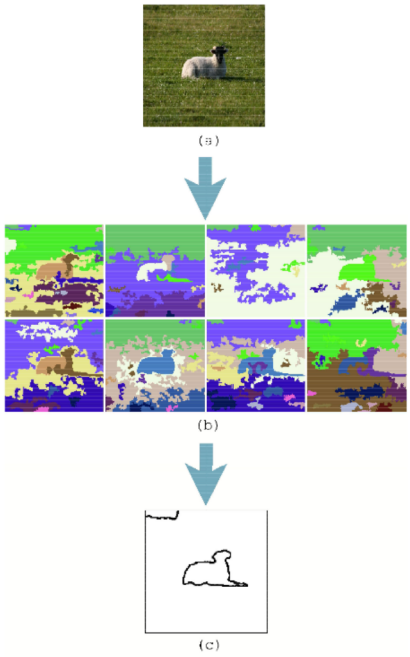

Fig. 2. Regions generation by the GCEBFM algorithm [32]. (a) Input image. (b) Examples of initial segmentation ensemble. (c) Segmentation result.

the region map or the segmentation of the input image (see Section III-A) and the region map of each image in the dataset. This new similarity criterion was recently proposed in the segmentation fusion framework [31] based on the *median partition* solution (which conceptually defines the consensus segmentation as being the partition that minimizes the average pairwise distance between itself and all other segmentations) and before that, as a quantitative metric to compare and evaluate a machine segmentation with multiple (possible) ground truths (i.e., manually segmented images provided by experts) [35]. Based on this metric, a perfect correspondence is yielded if each region in one of the segmentation is a subset or geometrically similar to a region in the other segmentation (this appealing property inherent to GCE makes this criterion relatively invariant to a possible over-segmentation). The GCE measure is originated from the so-called local refinement error (LRE) [35] which is expressed at each pixel. Mathematically, let n be the number of pixels within the image I and let $R_I = \{r_I^1, r_I^2, \ldots, r_I^{nb_I}\}$ & $R_{\mathrm{M}} = \{r_M^1, r_M^2, \ldots, r_M^{nb_M}\}$ be, respectively, the segmentation result of the input image to be measured and the segmentation of an image that belongs to the dataset, nb_I being the number of segments or regions in R_{I} and nb_M the number of regions in R_{M}. Let now p_i be a particular pixel and the couple $(r_{\mathrm{I}}^{<p_i>}, r_{\mathrm{M}}^{<p_i>})$ be the two segments including this pixel (respectively, in R_{I} and R_{M}). The local refinement error (LRE) can be computed at a pixel p_i as follows:

$$\mathrm{LRE}(r_{\mathrm{I}}, r_{\mathrm{M}}, p_i) = \frac{|r_{\mathrm{I}}^{<p_i>} \setminus r_{\mathrm{M}}^{<p_i>}|}{|r_{\mathrm{I}}^{<p_i>}|}. \quad (1)$$

Where $|r|$ denotes the cardinality of the set of pixels r and $\setminus$ represents the algebraic operator of difference. Particularly, a value of 1 means that the two regions overlap, in an inconsistent manner, on the contrary, an error of 0 expresses that the pixel is practically included in the refinement area [33]. A great way of forcing all local refinement to be in the same direction is to combine the LRE. On this basis, every pixel p_i must be computed twice, once in each sense, and in fact, gives as result the so-called global consistency error (GCE):

$$\mathrm{GCE}^{\star}(R_{\mathrm{I}}, R_{\mathrm{M}}) = \frac{1}{2n}\left\{\sum_{i=1}^{n}\mathrm{LRE}(r_{\mathrm{I}}, r_{\mathrm{M}}, p_i) + \sum_{i=1}^{n}\mathrm{LRE}(r_{\mathrm{M}}, r_{\mathrm{I}}, p_i)\right\}. \quad (2)$$

The GCE* value belongs in the interval of $[0, 1]$, on the one hand, a value of 0 expresses a maximum similarity between the two segmentations R_{I} and R_{M}, on the other hand, a value of 1 represents a bad match or correspondence between the two segmentations to be compared.

Finally, based on this GCE distance and in ascending order from the query image, we rank all the images of the entire dataset T. Then, we eliminate unhelpful images that have a higher GCE value, and we select a subset of images M from the entire dataset T as the retrieval set.

C. Region Features

A key idea with the proposed approach is that it simply uses large regions as the basic semantic unit. To perform the labeling process, we define the characteristics of those regions by extracting different features for each one. These used features are divided into five types;

- *Color:* This feature gives a relevant information about the statistical distribution of color related to each region. For each pixel, we estimate the re-quantized color histogram, with equidistant binning ($P_{BIN} = 5$) for each color channel (RGB), by considering the set of color values existing in an overlapping squared neighborhood ($SN = 7$) centered around this pixel. A normalized re-quantized color histogram is then estimated for each region by simply averaging the local histograms of each pixel belonging to the same region.
- *Texture:* To quantify the perceived texture of different regions in an image we use three features:
  - *Histogram of oriented gradients (HOG):* We compute the 40-bin normalized HOG with 4 different directions (respectively, vertical, horizontal, right diagonal, and left diagonal) and 10 amplitude values. By doing so, each histogram is computed on the luminance component of each pixel contained in an overlapping squared neighborhood ($SN = 7$) centered around each pixel in the image. Then, we average all histograms of pixels

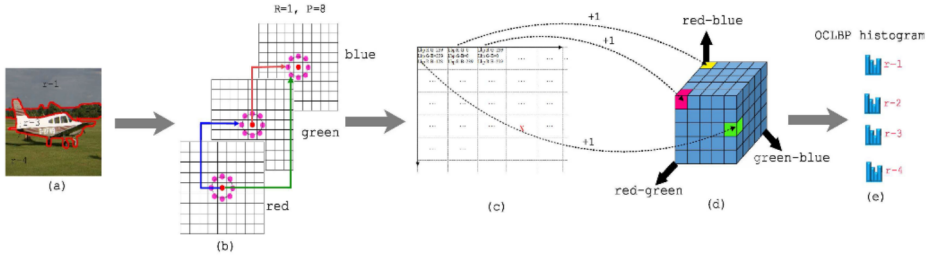

Fig. 3. Generation of the OCLBP histogram for each region. (a) The regions map of the input image. (b) Estimation of LBP value of a center pixel from one color channel based on neighborhoods from another channel [see (4)]. (c)-(d) Estimation, for each pixel X, of the N_b bin descriptor $q = 5$ in the cube of pair channels. Each $LbpR - G_X$, $LbpR - B_X$, $LbpG - B_X$ value associated with each pixel contained in a squared neighborhood region of size 7×7 centered at a pixel X, increments $(+1)$ a particular bin. (e) OCLBP histogram of each region.

which belong to the same region. Note that this region-based strategy of normalization aims to make this feature more invariant to changes in shading and illumination comparatively to a pixel-based approach.

– *Opponent color local binary pattern (OCLBP):* The original LBP operator proposed by Ojala *et al.* [36] was aimed to represent statistics of micro patterns contained in an image by encoding the difference between the pixel value of the center point and those of its neighbors. Formally, let I be a color image and let q_c be the value of the center pixel c of a local neighborhood and let $q_p(p = 0, ..., P-1)$ be the values of P equally spaced pixels on a circle of radius R that form a circularly symmetric set of neighbors. If the coordinates of q_c are $(0, 0)$, then the coordinates of q_p are given by $(R\sin(\frac{2\pi p}{P}), R\cos(\frac{2\pi p}{P}))$. Particularly, a bilinear interpolation is used to estimate the values of neighbors which do not fall exactly in the center of a pixel. The LBP operator on this pixel (c) is defined as follows:

$$LBP_{P,R} = \sum_{p=0}^{P-1} s(q_s - q_c)2^p, \; s(x) = \begin{cases} 1, & x \geq 0 \\ 0, & x < 0. \end{cases} \tag{3}$$

In our method we apply the opponent color version of LBP (OCLBP) presented in [37] and used recently in [38]. The idea within this extended version is to take a center pixel from one color channel and neighborhood from other color channel. For example, the OCLBP operator for a pixel c and between color channel pair (C_a, C_b) can be defined as:

$$OCLBP_{P,R}(C_a, C_b) = \sum_{p=0}^{P-1} s\left(q_s^{C_a} - q_c^{C_b}\right) 2^p. \tag{4}$$

After computing the OCLBP for three pairs of color channels (red-green, red-blue and green-blue), as input multidimensional descriptor of feature, we compute the set of values of the re-quantized OCLBP histogram (in each OCLBP result of color channel pair), with equidistant binning, $P_{BIN} = 5$. Thus, each histogram of 125 bins (as the feature descriptor) is estimated at an overlapping, fixed size squared ($N_w = 7$) neighborhood centered around the pixel. Finally, we average all histograms of pixels which belong to the same region (see Fig. 3).

Algorithm 1: Estimation of the Laplacian Operator

Mathematical notation:

r Radius (r = 1)

1: **for** each pixel $x(l, k)$ with color value R^x, G^x, B^x **do**
2: $x(l, k) = 1/3 \times (R^{x(l,k)} + G^{x(l,k)} + B^{x(l,k)})$
3: **end for**
4: **for** each pixel $x(l, k)$ **do**
5: $X_0(l, k) = \log(1 + x(l, k + r) - 2 \times x(l, k) + x(l, k - r))$
6: $X_1(l, k) = \log(1 + x(l + r, k) - 2 \times x(l, k) + x(l - r, k))$
7: $X_2(l, k) =$
$\log(1 + x(l, k + r) - 2 \times x(l, k) + x(l - r, k - r))$
8: **end for**

– *Laplacian operator (LAP):* In order to more efficiently capture local textural properties of each region, we also propose a new criterion derived from the Laplacian operator expressed in the logarithmic space [39] which efficiently complements the HOG features. The two steps of the estimation of this criterion are summarized in Algorithm 1.

- *Context:* As the context plays an important role in natural human recognition of objects and scene understanding [40], we decide to exploit the semantic contextual information around each region. More precisely, we compute the z-bin (z is the number of classes in the dataset) normalized histogram over the labels of the neighbors of each region excluding its own semantic label.
- *Shape:* Motivated by the efficacy of this classic feature, and in order to provide a simple geometric property, in our approach, we calculate the normalized area (i.e, the number of pixels in a region divided by the number of pixels within the image) of each region in the image.

- *Location:* This feature aims to capture the global position of each region with respect to the topmost pixel in the image (by computing the maximum y-coordinate). For example, sky region tends to have the minimum distance to the horizon.

D. Image Labeling

1) Principle: After extracting the feature descriptors used to describe regions and given 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.

More formally, let us assume that we have an input image I and its region segmentation $R_I = \{r_I^1, r_I^2, \ldots, r_I^m\}$ 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 labeling problem $\hat{S}_{\text{MC}} = \{s_I^1, s_I^2, \ldots, s_I^m\}$ is formulated as the result of the following multi-criteria optimization problem:

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

with:

$$\begin{aligned}
&\overline{\text{MC}}\left(I, R_I, S, \{\mathcal{I}_k, \mathcal{S}_k\}_{k \leq K}\right) \\
&= \alpha_1 \sum_{i=1}^{m} \text{COL}\left(I, r_I^i, s_I^i, \{\mathcal{I}_k, \mathcal{S}_k\}^{s_I^i}\right) \\
&+ \alpha_2 \sum_{i=1}^{m} \text{TEX}\left(I, r_I^i, s_I^i, \{\mathcal{I}_k, \mathcal{S}_k\}^{s_I^i}\right) \\
&+ \alpha_3 \sum_{i=1}^{m} \text{OCLBP}\left(I, r_I^i, s_I^i, \{\mathcal{I}_k, \mathcal{S}_k\}^{s_I^i}\right) \\
&+ \alpha_4 \sum_{i=1}^{m} \text{LAP}\left(I, r_I^i, s_I^i, \{\mathcal{I}_k, \mathcal{S}_k\}^{s_I^i}\right) \\
&+ \alpha_5 \sum_{i=1}^{m} \text{SHA}\left(r_I^i, s_I^i, \{\mathcal{I}_k, \mathcal{S}_k\}^{s_I^i}\right) \\
&+ \alpha_6 \sum_{i=1}^{m} \text{LOC}\left(r_I^i, s_I^i, \{\mathcal{I}_k, \mathcal{S}_k\}^{s_I^i}\right) \\
&+ \alpha_7 \sum_{i=1}^{m} \frac{1}{h} \left\{ \sum \text{CTX}\left(r_I^m, s_I^i, \{\mathcal{I}_k, \mathcal{S}_k\}^{s_I^i}\right) \right\}. \quad (5)
\end{aligned}$$

Where the parameters $\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5, \alpha_6$ and α_7 are used to weight the different terms of this energy function. COL, TEX, OCLBP, LAP, SHA, LOC and CTX designate respectively the different energy terms, or nonparametric 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, let $\{\mathcal{C}\}^{s_I^i} = \{\mathcal{I}_k, \mathcal{S}_k\}^{s_I^i}$ denotes the set of images $\mathcal{I}_k$ and their associated semantic segmentation solutions $\mathcal{S}_k$ (belonging to the training corpus) that contain a region semantically labeled s_I^i and let also h be the total number of those semantic segmentations in the corpus $\{\mathcal{C}\}^{s_I^i}$ (see Table I). Then, COL(.), TEX(.), OCLBP(.), LAP(.) and CTX(.) are, respectively, the minimum Ruzicka distance[2] between the p-bin normalized color histogram, the q-bin normalized histogram of oriented gradients (HOG), the p-bin normalized OCLBP histogram, the p-bin normalized LAP histogram,the z-bin normalized histogram of semantic labels of r_I^i and those of each region corresponding to the semantic label assigned to r_I^i (i.e., s_I^i) and existing in $\{\mathcal{C}\}^{s_I^i}$. Also, LOC(.) and SHA(.) are, respectively, the minimum absolute distance between the normalized area, the height of the topmost pixel existing in the region r_I^i, and normalized area 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 $\{\mathcal{C}\}^{s_I^i}$.

TABLE I
SUMMARY OF THE COMBINED CRITERIA USED IN OUR MODEL

TYPE	CRITERION	DIMENSION
Color	Color histogram	125
Texture	Oriented gradient histogram	40
	Opponent color local binary pattern histogram	125
	Laplacian operator histogram	125
Shape	Pixel area	1
Location	Top height	1
Context	Context histogram	21

2) Optimization of the Energy Function: The proposed semantic segmentation model of multiple label fields is formulated as a global optimization problem incorporating a nonlinear multi-objective function. In order to achieve the minimum of this energy function [see (5)], approximation approaches based on different optimization algorithms such as the exploration/selection/estimation (ESE) [41], 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 avoid this problem, in this work we adopt the iterated conditional modes (ICM) method proposed by Besag [42] (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 100 iterations according to our experiments). The entire pseudo-code of our MC-SSM based on ICM is presented in Algorithm 2.

IV. EXPERIMENTS

A. Datasets

To evaluate the performance of our model, we compared it with different nonparametric methods, tested on two challenging semantic segmentation datasets; Microsoft Research Cambridge dataset [34] and the Stanford background dataset [43].

1) Microsoft Research Cambridge Dataset (MSRC-21): The MSRC-21 (v2) dataset[3] is an extension of the MSRC-9 (v1)

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

[3] The MSRC-21 dataset can be downloaded here: http://www.cs.cmu.edu/ tmalisie/projects/bmvc07/.

Algorithm 2: MC-Semantic Segmentation Model algorithm

Mathematical notation:

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

Input: $I, \{\mathcal{I}_k\}_{k\leq K}, \{S_k\}_{k\leq K}$
Output: $\hat{S}_{MC}$

A. Initialization:
1: Segment image I into different coherent regions R_I (with the GCEBFM algorithm)
2: Assign class label for each r_i region $\in R_I$ using random element from $\mathcal{E}$
B. Steepest Local Energy Descent:
3: **while** $p < T_{\max}$ **do**
4: **for** for each r_i region $\in R_I$ **do**
5: Draw a new label y according to the uniform distribution in the set $\mathcal{E}$
6: Let $R_I^{[p],\textbf{new}}$ the new semantic segmentation map including r_i with the class label y
7: Compute $\overline{\text{MC}}\,(I, R_I^{[p],\textbf{new}}, S, \{\mathcal{I}_k, S_k\}_{k\leq K})$ [see (5)]
8: **if** $\overline{\text{MC}}\,(I, R_I^{[p],\textbf{new}}, S, \{\mathcal{I}_k, S_k\}_{k\leq K}) < \overline{\text{MC}}\,(I, R_I^{[p]}, S, \{\mathcal{I}_k, S_k\}_{k\leq K})$ **then**
9: $\overline{\text{MC}} = \overline{\text{MC}}^{\textbf{new}}$
10: $R_I^{[p]} = R_I^{[p],\textbf{new}}$
11: $\hat{S}_{MC} = R_I^{[p]}$
12: **end if**
13: **end for**
14: $p \leftarrow p+1$
15: **end while**

dataset. It contains 591 color images with corresponding ground truth labeling for 23 object classes (building, grass, tree, cow, etc.). Among the 23 object classes, only 21 classes are commonly used. The unused labels are (void = 0, horse = 5, mountain = 8) due to background or too few training samples.

2) Stanford Background Dataset (SBD): The SBD dataset[4] contains a set of outdoor scene images imported from existing public datasets: LabelMe [44], MSRC [34], PASCAL VOC [45] and Geometric Context [46]. Each image in this dataset contains at least one foreground object. The dataset is pixel-wise annotated (horizon location, pixel semantic class, pixel geometric class and image region) for evaluating methods for semantic scene understanding.

B. Evaluation Metrics

To provide a basis of comparison for the MC-SSM model, we quantitatively evaluate the annotation performance from two levels, which are widely used for evaluating the performances of related tasks. The first is the global (overall) per-pixel accuracy (GPA) which represents the total proportion of pixels correctly labeled. Mathematically, the global accuracy is computed as:

$$\text{GPA} = \frac{\sum_{i=1}^{n} \mathrm{v}(x)}{n}, \quad \mathrm{v}(x)\begin{cases}1 & y_i = l_i. \\ 0 & \text{otherwise}\end{cases} \tag{6}$$

Where $v(.)$ denotes the indicator function, n is the number of pixels within the input image, y_i represents the label for pixel i predicted by the algorithm and l_i denotes the ground truth label for pixel i. The second level is the average per-class accuracy (ACA) which represents the average proportion of pixels correctly labeled in each category. Formally, the class-averaged accuracy is computed as follows:

$$\text{ACA} = \frac{1}{|C|}\sum_{c\in C}\frac{\sum_{i=1}^{n\times nb}\mathrm{v}(y_i = l_i) \wedge \mathrm{v}(l_i = c)}{\sum_{i=1}^{n\times nb}\mathrm{v}(l_i = c)}. \tag{7}$$

Where |C| denotes the number of classes within the input image, nb is the number of images in the dataset and $\wedge$ represents the logic operator *And*.

C. Results and Discussion

To validate our model on the MSRC-21 dataset, 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.

TABLE II
PERFORMANCE OF OUR MODEL ON THE MSRC-21 SEGMENTATION DATASET IN TERMS OF GLOBAL PER-PIXEL ACCURACY AND AVERAGE PER-CLASS ACCURACY (HIGHER IS BETTER)

ALGORITHMS	PERFORMANCE MEASURES	
	Global (GPA)	Average (ACA)
Nonparametric (non-learning-based) methods		
MC-SSM	0.75	0.63
SuperParsing [51] in [54]	0.62	NA
Hierarchical [53]	NA	0.53
Parametric (learning-based) methods		
SVM on segment [8]	0.51	NA
CRF on segment [8]	0.64	NA
CRF+N=2 [55] in [8]	0.68	NA
CRF+N=3 [55] in [8]	0.68	NA
SVM on region [8]	0.69	NA
Tree model [8]	0.70	NA
TextonBoost [34]	0.72	0.58
Graphical model [56]	0.75	0.65
Auto-context [49]	0.75	NA
GP [57]	0.72	NA
SVMs in [58]	0.64	0.47
AdaBoost in [58]	0.69	0.52
SSVMs in [58]	0.71	0.57
CRFTree in [58]	0.74	0.65
CRFTree (FL) [58]	0.82	0.75
Csurka [60] in [59]	NA	0.63

[4]The SBD dataset is publicly accessible via this link: http://dags.stanford.edu/data/iccv09Data.tar.gz

TABLE III
ACCURACY OF SEGMENTATION FOR THE MSRC 21-CLASS DATASET. CONFUSION MATRIX WITH PERCENTAGES ROW-NORMALIZED. THE OVERALL PER-PIXEL ACCURACY IS 75%

TRUE CLASS \ INFERRED CLASS	building	grass	tree	cow	sheep	sky	aeroplane	water	face	car	bicycle	flower	sign	bird	book	chair	road	cat	dog	body	boat
building	**53.6**	3.6	3.6		3.6	1.2	2.4	3.6		2.4			1.2	9.5			3.6	7.1	2.4	1.2	1.2
grass		**89.9**	2.9		0.7			0.7						0.7			3.6	0.7	0.7		
tree	5.6	12.5	**55.6**	2.8	2.8		1.4							12.5			1.4	4.2			1.4
cow				**72.7**	9.1									9.1					9.1		
sheep				5.0	**80.0**									5.0				5.0	5.0		
sky						**95.1**		2.4						1.2					1.2		
aeroplane	6.2						**87.5**			6.2											
water		2.5	2.5			10.0		**57.5**						2.5			25.0				
face									**69.7**	3.0				9.1			3.0		6.1	6.1	3.0
car	8.3						8.3			**58.3**											25.0
bicycle	11.8										**64.7**			11.8				11.8			
flower		5.6		5.6						5.6		**61.1**	5.6		5.6		5.6	5.6			
sign								5.6	5.6				**72.2**	5.6		5.6	5.6				
bird			5.0		10.0								5.0	**50.0**		5.0		15.0	5.0	5.0	
book			5.6		11.1										**83.3**						
chair	11.8			11.8	11.8							5.9		17.6		**17.6**		17.6	5.9		
road	5.7	2.3						8.0				1.1		1.1		1.1	**72.4**	2.3	3.4	1.1	1.1
cat				7.7	7.7										7.7			**76.9**			
dog	6.2			12.5	18.8									18.8	6.2			12.5	**18.8**	6.2	
body	2.7			2.7	2.7		2.7		2.7				2.7	24.3			2.7	5.4		**48.6**	2.7
boat	23.5						5.9			5.9			5.9					17.6	5.9		**35.3**

TABLE IV
TEXTURE SIMILARITY LEVEL BETWEEN THE DIFFERENT CLASSES OF THE MSRC 21-CLASS DATASET. LOWER VALUES INDICATE MORE SIMILARITY

CLASS \ CLASS	building	grass	tree	cow	sheep	sky	aeroplane	water	face	car	bicycle	flower	sign	bird	book	chair	road	cat	dog	body	boat
building		19.2	17.9	15.7	15.6	37.6	15.4	23.0	17.9	16.5	20.1	18.0	22.2	16.8	18.6	15.7	19.2	16.4	16.2	17.6	16.7
grass			21.2	17.5	18.1	39.8	17.4	22.0	20.3	19.5	25.5	19.9	25.8	19.9	22.7	19.0	17.9	18.7	17.1	19.3	19.9
tree				18.0	15.2	42.2	17.1	26.0	16.2	17.4	13.0	19.5	26.5	16.0	22.0	14.6	22.0	17.1	18.4	19.9	15.4
cow					12.3	37.3	16.7	22.2	15.4	17.2	21.3	15.5	22.1	15.6	16.2	15.0	17.3	14.2	12.4	15.4	17.6
sheep						39.5	16.6	23.5	13.8	16.7	16.3	16.1	23.9	14.0	17.6	13.1	18.4	13.6	12.9	16.0	16.2
sky							38.1	36.3	41.8	38.1	45.4	38.6	30.1	40.1	34.5	40.7	37.9	39.1	38.3	37.2	40.9
aeroplane								19.8	18.3	11.6	19.9	17.7	20.5	15.8	20.6	14.2	16.6	15.9	16.3	17.1	12.0
water									25.4	22.2	30.1	23.7	25.4	24.3	25.0	23.6	21.3	23.2	22.0	22.7	23.3
face										18.5	16.0	18.2	26.1	16.2	20.4	15.1	20.9	16.3	16.0	18.6	17.1
car											19.1	18.6	21.3	16.6	20.4	15.2	18.8	16.8	17.4	18.2	13.9
bicycle												21.7	29.1	15.9	23.8	14.3	26.2	18.6	22.0	22.7	15.6
flower													23.5	17.9	18.6	17.2	19.4	17.2	16.1	17.9	18.7
sign														23.9	20.5	23.8	23.9	23.2	22.4	22.6	23.6
bird															19.7	14.3	20.0	15.9	16.3	18.0	15.5
book																19.4	21.6	18.6	17.4	18.4	22.0
chair																	19.4	14.7	15.5	17.2	13.9
road																		19.1	17.4	19.1	20.3
cat																			14.6	16.9	16.0
dog																				15.7	17.4
body																					18.8
boat																					

TABLE V
COLOR SIMILARITY LEVEL BETWEEN THE DIFFERENT CLASSES OF THE MSRC 21-CLASS DATASET. LOWER VALUES INDICATE MORE SIMILARITY

CLASS	building	grass	tree	cow	sheep	sky	aeroplane	water	face	car	bicycle	flower	sign	bird	book	chair	road	cat	dog	body	boat
building		90.9	83.1	84.5	72.7	93.5	71.9	85.0	84.4	79.0	70.0	91.8	85.6	76.5	81.2	77.1	81.0	75.3	77.4	83.0	76.3
grass			85.4	91.0	88.4	99.0	89.7	95.6	94.0	93.8	88.4	95.1	96.1	89.9	92.6	90.1	94.3	91.4	91.1	94.2	93.5
tree				77.4	76.8	97.9	74.9	91.3	87.6	76.6	72.7	91.2	89.6	78.4	82.8	82.4	90.9	73.8	79.0	84.9	82.4
cow					79.0	96.5	76.6	91.8	82.1	77.4	75.9	89.8	88.3	80.0	80.4	80.9	90.2	72.6	77.3	83.5	82.3
sheep						92.7	66.1	84.3	79.9	74.0	64.2	88.2	84.0	72.0	76.9	73.6	80.0	69.2	72.6	78.8	73.7
sky							91.6	93.3	97.3	95.2	96.5	96.7	93.6	91.5	97.2	96.1	91.0	95.1	94.7	94.7	93.5
aeroplane								82.1	80.6	68.4	62.3	86.9	80.2	68.8	75.4	72.4	78.4	64.8	71.1	77.8	69.7
water									93.2	86.4	83.3	95.4	88.4	86.2	90.0	87.4	85.5	86.7	87.7	90.2	83.7
face										85.0	81.8	91.4	90.7	84.2	83.2	81.6	88.4	81.8	79.9	83.7	85.4
car											65.8	89.5	83.3	75.3	76.1	79.6	85.9	68.1	76.5	78.8	71.4
bicycle												87.3	81.6	68.4	71.6	70.6	79.9	64.0	70.3	76.2	66.9
flower													91.9	88.8	88.5	89.1	92.2	88.2	89.1	90.4	89.5
sign														84.2	86.2	85.8	85.1	82.5	84.8	87.2	82.9
bird															79.8	77.0	81.9	72.1	75.0	80.7	75.4
book																78.8	86.2	75.6	78.7	81.4	77.7
chair																	82.3	74.9	75.7	82.1	77.0
road																		85.0	85.0	88.3	83.9
cat																			69.8	76.7	71.3
dog																				80.0	76.0
body																					81.1
boat																					

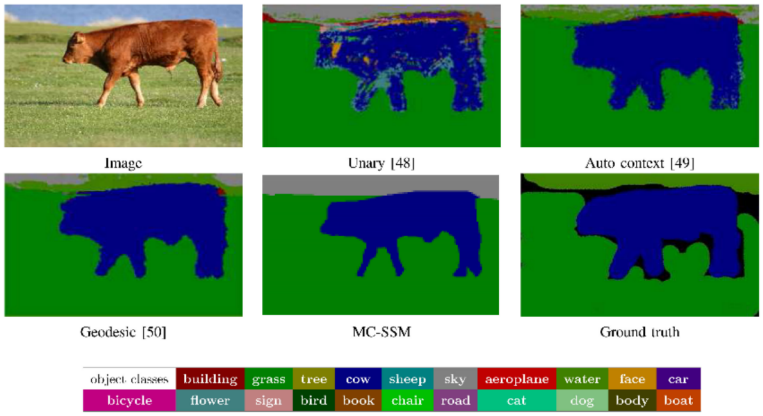

Fig. 4. Example of segmentation result obtained by our algorithm MC-SSM on an input image from the MSRC-21 compared to other algorithms.

To guarantee the integrity of the benchmark results, the seven weight parameters of our algorithm [i.e., α_1, α_2, α_3, α_4, α_5, α_6 and α_7, see (5)] are optimized on the ensemble of 276 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 $\alpha_1 = 1.83, \alpha_2 = 1.53, \alpha_3 = 1.55, \alpha_4 = 1.44, \alpha_5 = 1.35, \alpha_6 = 1.70$ and $\alpha_7 = 1$, are reliable hyper-parameters for the model yielding the best performance.

As we show in Table II, MC-SSM outperforms the nonparametric SuperParsing method [51] with a GPA and ACA scores equal to, respectively, 0.75 and 0.63 (we perform tests on the 315 test images). Also, compared with state-of-the-art parametric methods, our method gives good results while not requiring

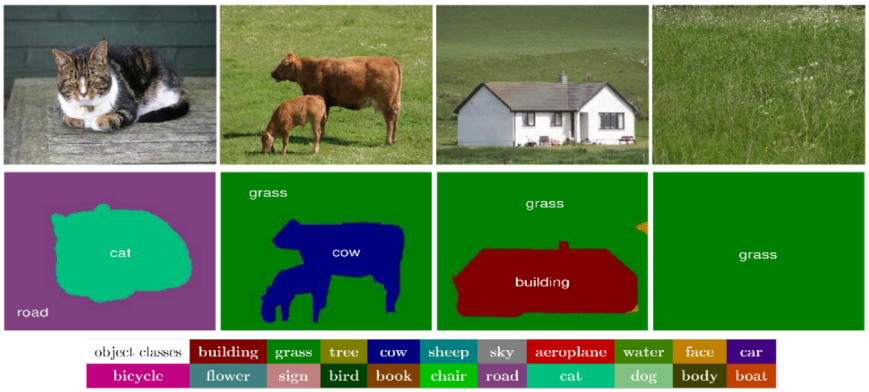

Fig. 5. Example results obtained by our MC-SSM model on the MSRC-21 dataset (for more clarity, we have superimposed textual labels on the resulting segmentations).

TABLE VI
PERFORMANCE OF OUR MODEL ON THE STANFORD BACKGROUND DATASET (SBD) IN TERMS OF GLOBAL PER-PIXEL ACCURACY AND AVERAGE PER-CLASS ACCURACY (HIGHER IS BETTER)

ALGORITHMS	PERFORMANCE MEASURES	
	Global (GPA)	Average (ACA)
Nonparametric (non-learning-based) methods		
MC-SSM	0.68	0.62
SuperParsing [51]	0.76	NA
Parametric (learning-based) methods		
SVM on segment [8]	0.51	NA
ContextL [52]	0.54	0.45
CRF on segment [8]	0.62	NA
CRF+N=2 [55] in [8]	0.67	NA
CRF+N=3 [55] in [8]	0.66	NA
Singlescale ConvNet [7]	0.66	0.57
SVM on region [8]	0.69	NA
Tree model [8]	0.69	NA
ConvNet [52]	0.69	0.66
Leaf Level [61]	0.73	0.58
DeepLab-largeFOV [62] in [63]	NA	0.65
RWN-largeFOV [63]	NA	0.68
Recurrent CNNs [64]	0.76	0.67
HGDN [65]	0.82	0.72

TABLE VII
ACCURACY OF SEGMENTATION FOR THE SBD DATASET. CONFUSION MATRIX WITH PERCENTAGES ROW-NORMALIZED. THE OVERALL PER-PIXEL ACCURACY IS 68%

		INFERRED CLASS							
		sky	tree	road	grass	water	building	mountain	foreground
TRUE CLASS	sky	**92.3**	2.5	2.7		1.5	0.2	0.2	0.7
	tree	0.4	**32.1**	2.7	1.6	0.4	5.7	1.8	55.4
	road	1.3	2.0	**80.2**	1.4	8.1	1.8	1.8	3.4
	grass		9.8	9.3	**52.1**	2.1	18.0	4.6	4.1
	water	5.2	2.1	35.1	1.0	**43.3**	8.2	3.1	2.1
	building	0.4	12.5	3.7	0.7	0.4	**74.1**	1.1	7.1
	mountain	4.2	35.2	16.9	2.8	4.2	12.7	**15.5**	8.5
	foreground	0.6	33.6	3.9	1.1	2.0	15.5	5.0	**38.4**

expensive model training and being much simpler. It is worth mentioning that parametric scene parsing methods have a small advantage in accuracy over nonparametric methods. However, they require large amounts of model training, making them less practical for open datasets [47]. The confusion matrix experimented from the MSRC-21 dataset is shown in Table III. From this table we can see that better result in terms of class accuracy is yielded for the following classes; *sky*, *grass*, *aeroplane*, *sheep* and *book*, with values are higher than 80%. However, lower accuracy is achieved for the *chair* class with a value equal to 17.6%, this class is often confused with the *bird* class due to the similarity in color and texture between these two classes (see Table IV and Table V). Additionally, we present a qualitative comparison with other methods; Unary [48], Auto context [49] and Geodesic [50] (see Fig. 4). Also, Fig. 5 shows example results of success on the MSRC-21 generated by our algorithm.

In addition, we validated our model on the SBD dataset and we adopt the same evaluation strategy, the leave-one-out, but for the entire dataset as we used the same value of the parameters fixed on the training set of the MSRC-21 dataset. Table VI shows that our model is still competitive with different methods with a GPA value equal to 0.68 and ACA value equal to 0.53. These values are less better compared to those achieved on the

TABLE VIII
PERFORMANCE OF OUR MODEL USING SINGLE AND MULTIPLE CRITERIA (ON THE MSRC-21 DATASET)

	CRITERION	MEASURES (GPA)	MEASURES (ACA)
SINGLE CRITERION	CTX	0.13	0.07
	LOC	0.15	0.13
	SHA	0.19	0.13
	TEX	0.26	0.18
	OCLBP	0.54	0.43
	LAP	0.59	0.49
	COL	0.65	0.55
MULTIPLE CRITERIA	TEX+CTX	0.27	0.19
	TEX+CTX+LOC+SHA	0.38	0.23
	TEX+CTX+LOC+SHA+COL	0.71	0.58
	TEX+CTX+LOC+SHA+COL+OCLBP+LAP	**0.75**	**0.63**

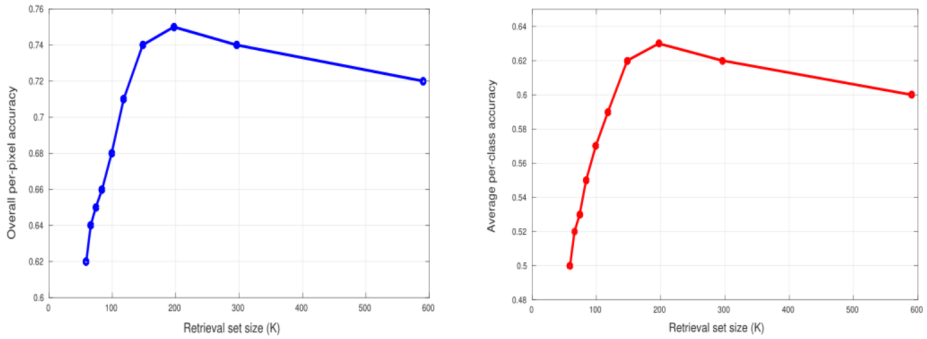

Fig. 6. Effects of varying the retrieval set size K for the MSRC-21 dataset; shown are the overall per-pixel accuracy and the average per-class accuracy.

MSRC-21 dataset. This result is not surprising, because the SBD dataset contains a foreground class that refers to different types of objects which increases significantly the intra-class variability.

Table VII shows the confusion matrix for our model in the SBD dataset. From this table, we can note that better result in terms of class accuracy is yielded for the following classes; *sky* and *grass* classes, with values are higher than 80%. In contrast, lower accuracy is achieved for the *mountain* class with a value equal to 15.5%.

We have also tested the effects of varying the retrieval set size K in Fig. 6. This test shows that $K = 197$ (the 1/3 of the dataset) is a reliable value that yielding the best accuracy for our model. As another evaluation test, in Table VII we report the results of our model using single criterion and multiple criteria. In fact, compared to the mono-criterion case our multi-criteria approach (in bold) achieves a better result. This shows clearly that our strategy of combining different criteria is effective. In addition, in this table, we present the relative importance of each used feature. As we can see, color histogram, OCLBP and Laplacian operator histogram are the criteria that provide the best accuracy scores. In order to test the convergence properties of our iterative optimization procedure, we have tested our algorithm with different random initializations (step 2 in Algorithm 2) and we have found similar results, this result shows clearly that the consensus cost function [see Eq. (5)] is nearly convex. This also means that the proposed semantic labeling model is numerically rendered well-posed (and the optimization problem tractable) thanks to appropriate convex constraints or appropriate feature descriptors for this kind of problem. Also, we have evaluated the proposed model with different iteration numbers of the optimization algorithm and we have found that $T_{\max} = 100$ is the best value which gives the asymptotic result in terms of GPA and ACA on the MSRC-21 dataset (see Fig. 7).

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

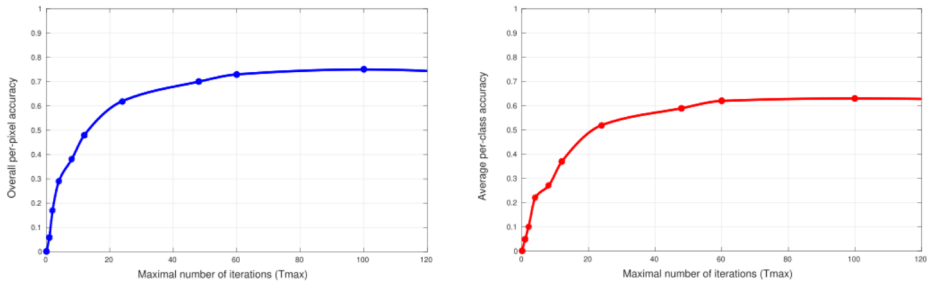

Fig. 7. Evolution of the overall per-pixel accuracy and the average global per-class accuracy along the number of iterations of the proposed MC-SSM starting from a random initialization on the MSRC-21 dataset.

TABLE IX
COMPARISON OF TIME COMPETITION

ALGORITHMS	SEGMENTATION TIME	IMAGE SIZE
SuperParsing [51]	NA (train.) + 6 minutes (test.)	640 × 480
TextonBoost [34]	42 hours (train.) + 3 minutes (test.)	320 × 213
Auto-Context [49]	NA (train.) + 70 seconds (test.)	300 × 200
MC-SSM	5 minutes	240 × 240

(line segment detector) descriptors or other similarity measures between segmentations.

D. Computation Time

The computational complexity of the proposed model depends on two factors; the number of the images in the dataset and the number of the used criteria (combined as a global energy function). On the MSRC-21 dataset, the execution time takes, on average, between 5 and 6 minutes for an Intel 64 Processor core i7-4800MQ, 2.7 GHz, 8 GB of RAM memory and non-optimized code running on Linux for a 240×240 image. More accurately, the labeling process takes 0.14 second and the geometric retrieval step takes 0.32 second. However, the computation time of the proposed model (for each image) is mainly occupied by the region generation code with 205 seconds and the features extraction (from the full dataset) with 171 seconds. The former can be reduced by a parallelized implementation while the latter can be easily sped up by performing the extraction only once and then storing the extracted features on a data structure. We summarize the available segmentation time required by other related works in Table IX. The whole unoptimized and unparallelized implementation of our method was developed using the C++ language.

V. CONCLUSION

The aim of this present research was to address the problem of scene parsing (also called semantic segmentation). Towards this goal, we proposed a novel and simple energy-minimization based approach called the multi-criteria semantic segmentation model (MC-SSM).

Moreover, by using a new geometric retrieval strategy, we selected nearest neighbors from a database containing fully segmented and annotated images. This strategy is based on a new criterion called global consistency error (GCE). This criterion aims to find matches between the region map or the segmentation of the input image and the region map of each image in the dataset. In addition, the proposed cost function of this model combines efficiently different global nonparametric semantic likelihood energy terms computed from the (pre-)segmented regions of the (query) image and defined according to their structural properties (location, texture, color, context and shape). Furthermore, by combining different features into the energy or the objective function, our model integrates more information about the object possibly present in the scene. To optimize our energy-based model we resort to a simple and local optimization procedure derived from the iterative conditional modes (ICM) algorithm. Our approach achieved state-of-the-art performance in two popular datasets (MSRC-21 and SBD). An interesting finding that can be observed from the experiments, is that combining different criteria improves significantly the final result of scene parsing. This suggests our method to be a suitable alternative, to methods that require large datasets to train high-accuracy classifiers. Thus, the proposed model is dedicated to small datasets which are characterized by a limited number of available images. One area of future work will be, to improve further the classification accuracy by incorporating other criteria (possibly at different geometric and semantic abstraction levels).

REFERENCES

[1] F. Meng, H. Li, Q. Wu, K. N. Ngan, and J. Cai, "Seeds-based part segmentation by seeds propagation and region convexity decomposition," *IEEE Trans. Multimedia*, vol. 20, no. 2, pp. 310–322, Feb. 2018.
[2] S. P. Narote, P. N. Bhujbal, A. S. Narote, and D. M. Dhane, "A review of recent advances in lane detection and departure warning system,"*Pattern Recognit.*, vol. 76, pp. 216–234, 2018.
[3] K. Li, W. Tao, X. Liu, and L. Liu, "Iterative image segmentation with feature driven heuristic four-color labeling," *Pattern Recognit.*, vol. 76, pp. 69–79, 2018.

[4] K. Kim and S. W. Jung, "Interactive image segmentation using semi-transparent wearable glasses," *IEEE Trans. Multimedia*, vol. 20, no. 1, pp. 208–223, Jan. 2018.

[5] E. L. Andrade, J. C. Woods, E. Khan, and M. Ghanbari, "Region-based analysis and retrieval for tracking of semantic objects and provision of augmented information in interactive sport scenes," *IEEE Trans. Multimedia*, vol. 7, no. 6, pp. 1084–1096, Dec. 2005.

[6] O. Gupta, D. Raviv, and R. Raskar, "Illumination invariants in deep video expression recognition," *Pattern Recognit.*, vol. 76, pp. 25–35, 2018.

[7] C. Farabet, C. Couprie, L. Najman, and Y. LeCun, "Learning hierarchical features for scene labeling," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1915–1929, Aug. 2013.

[8] J. Xie, L. Yu , L. Zhu, and X. Chen, "Semantic image segmentation method with multiple adjacency trees and multiscale features," *Cogn. Comput.*, vol. 9, no. 2, pp. 168–179, 2017.

[9] G. Lin, A. Milan, C. Shen, and I. Reid, "RefineNet: Multi-path refinement networks for high-resolution semantic segmentation," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1925–1934.

[10] S. Gould, J. Rodgers, D. Cohen, G. Elidan, and D. Koller, " Multi-class segmentation with relative location prior," *Int. J. Comput. Vis.*, vol. 80, no. 3, pp. 300–316, 2008.

[11] R. Achanta *et al.*, "SLIC super-pixels compared to state-of-the-art superpixel methods," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2274–2282, Nov. 2012.

[12] B. Tung and J. J. Little, "Scene parsing by nonparametric label transfer of content-adaptive windows," *Comput. Vis. Image Understand*, vol. 143, pp. 191–200, 2016.

[13] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.

[14] E. Shelhamer, J. Long, C. Fowlkes, and D. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 640–651, Apr. 2017.

[15] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Region-based convolutional networks for accurate object detection and segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 1, pp. 142–158, Jan. 2016.

[16] Y. Li, H. Qi, J. Dai, X. Ji, and Y. Wei, "Fully convolutional instance-aware semantic segmentation," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4438–4446.

[17] B. Shuai, Z. Zuo, G. Wang, and B. Wang, "Scene parsing with integration of parametric and non-parametric models," *IEEE Trans. Image Process.*, vol. 25, no. 5, pp. 2379–2391, May 2016.

[18] S. Liu *et al.*, "Matching-CNN meets KNN: Quasi-Parametric human parsing," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Boston, MA, USA, Jun. 2015, pp. 1419–1427.

[19] X. An, S. Li, H. Qin, and A. Hao, "Automatic non-parametric image parsing via hierarchical semantic voting based on sparse-dense reconstruction and spatial-contextual cues," *Neurocomputing*, vol. 201, pp. 92–103, 2016.

[20] A. Khoreva, R. Benenson, J. Hosang, M. Hein, and B. Schiele, "Simple does it: Weakly supervised instance and semantic segmentation," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 376–385.

[21] J. Shotton, J. Winn, C. Rother, and A. Criminisi, "Texton- boost 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.

[22] J. Shotton and P. Kohli, "Semantic image segmentation," *in Computer Vision: A Reference Guide*, K. Ikeuchi, Ed. Berlin, Germany: Springer, 2014, pp. 713–716.

[23] C. Liu, J. Yuen, and A. Torralba, "Nonparametric scene parsing via label transfer," in *Dense Image Correspondences for Computer Vision*, T. Hassner and C. Liu, Eds. Berlin, Germany: Springer, 2016, pp. 207–236.

[24] C. Liu, J. Yuen, and A. Torralba, "SIFT flow: Dense correspondence across scenes and its applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 5, pp. 978–994, May 2011.

[25] Z. Liu, X. Li, P. Luo, C. C. Loy, and X. Tang, "Semantic image segmentation via deep parsing network," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1377–1385.

[26] 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.

[27] J. Tighe and S. Lazebnik, "Superparsing: Scalable nonparametric image parsing with superpixels," *Int. J. Comput. Vis.*, vol. 101, no. 2, pp. 329–349, 2013.

[28] M. Zand, S. Doraisamy, A. A. Halin, and M. R. Mustaffa, "Ontology-based semantic image segmentation using mixture models and multiple CRFs," *IEEE Trans. Image Process.*, vol. 25, no. 7, pp. 3233–3248, Jul. 2016.

[29] H. Zhang, T. Fang, X. Chen, Q. Zhao, and L. Quan, "Partial similarity based nonparametric scene parsing in certain environment," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2011, pp. 2241–2248.

[30] L. khelifi and M. Mignotte, "Semantic image segmentation using the ICM algorithm," in *Proc. 24th IEEE Int. Conf. Image Process.*, 2017, pp. 3080–3084.

[31] L. Khelifi and M. Mignotte, "A novel fusion approach based on the global consistency criterion to fusing multiple segmentations," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 47, no. 9, pp. 2489–2502, Sep. 2017.

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

[33] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *Proc. 8th Int. Conf. Comput. Vis.*, Jul. 2001, vol. 2, pp. 416–423.

[34] J. Shotton, J. Winn, C. Rother, and A. Criminisi, "TextonBoost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation," in *Proc. 9th Eur. Conf. Comput. Vis.,* 2006, pp. 1–15.

[35] A. Y. Yang, J. Wright, S. Sastry, and Y. Ma, "Unsupervised segmentation of natural images via lossy data compression," *Comput. Vis. Image Understand.*, vol. 110, no. 2, pp. 212–225, May 2008.

[36] T. Ojala, M. Pietikäinen, and T. Mäenpää, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, Jul. 2002.

[37] T. Maenpaa, M. Pietikainen, and J. Viertola, "Separating color and pattern information for color texture discrimination," in *Proc. 16th IEEE Int. Conf. Pattern Recognit.*, 2002, pp. 668–671.

[38] A. Joshi and A. K. Gangwar, "Color local phase quantization (CLPQ)- A new face representation approach using color texture cues," in *Proc. IEEE Int. Conf. Biometrics*, 2015, pp. 177–184.

[39] F. Y. Shih, "Mathematical preliminaries,"*Image Processing Pattern Recognition: Fundamentals Techniques*, Hoboken, NJ, USA: Wiley, 2010, pp. 17–38.

[40] S. Gould and X. He, "Scene understanding by labeling pixels," *Commun. ACM*, vol. 57, no. 11, pp. 68–77, 2014.

[41] 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, Aug. 2005.

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

[43] S. Gould, R. Fulton, and D. Koller, "Decomposing a scene into geometric and semantically consistent regions," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2009, pp. 1–8.

[44] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman, "LabelMe: A database and web-based tool for image annotation," *Int. J. Comput. Vis.*, vol. 77, no. 1, pp. 157–173, 2008.

[45] M. Everingham, L. Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, 2010.

[46] D. Hoiem, A. A. Efros, and M. Hebert, "Recovering surface layout from an image," *Int. J. Comput. Vis.*, vol. 75, no. 1, pp. 151–172, 2007.

[47] F. Tung *et al.*, "CollageParsing: Nonparametric scene parsing by adaptive overlapping windows," in *Proc. 13th Eur. Conf. Comput. Vis.*, 2014, pp. 511–525.

[48] 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, Jun. 2014.

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

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

[51] J. Tighe and S. Lazebnik, "Superparsing—Scalable nonparametric image parsing with superpixels," *Int. J. Comput. Vis.*, vol. 101, no. 2, pp. 329–349, 2013.

[52] T. Kekec, R. Emonet, È. Fromont, A. Trémeau, and C. Wolf, "Contextually constrained deep networks for scene labeling," in *Proc. Brit. Mach. Vis. Conf.*, 2014, pp. 1–11.
[53] D. Yuan and J. Qiang, "Hierarchical image segmentation using semantic edge constraint," in *Proc. IEEE Int. Conf. Real-Time Comput. Robot.*, 2016, pp. 82–87.
[54] A. Bassiouny and M. El-Saban, "Semantic segmentation as image representation for scene recognition," in *Proc. IEEE Int. Conf. Image Process.*, 2014, pp. 981–985.
[55] B. Fulkerson, A. Vedaldi, and S. Soatto, "Class segmentation and object localization with superpixel neighborhoods," in *Proc. 12th IEEE Int. Conf. Comput. Vis.*, 2009, pp. 670–677.
[56] L. Zhang and Q. Ji, "Image segmentation with a unified graphical model," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 8, pp. 1406–1425, Aug. 2010.
[57] Q. Li *et al.*, "Geodesic propagation for semantic labeling," *IEEE Trans. Image Process.*, vol. 23, no. 11, pp. 4812–4825, Nov. 2014.
[58] F. Liu, G. Lin, R. Qiao, and C. Shen, "Structured learning of tree potentials in CRF for image segmentation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 6, pp. 2631–2637, Jun. 2018.
[59] Y. Li, Y. Guo, Y. Kao, and R. He, "Image piece learning for weakly supervised semantic segmentation," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 47, no. 4, pp. 648–659, Apr. 2017.
[60] G. Csurka and F. Perronnin, "An efficient approach to semantic segmentation," *Int. J. Comput. Vis.*, vol. 95, no. 4, pp. 198–212, 2011.
[61] D. Munoz, J. A. Bagnell, and M. Hebert, "Stacked hierarchical labeling," in *Proc. Eur. Conf. Comp. Vis.*, Crete, Greece, 2010, pp. 57–70.
[62] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected CRFs," in *Proc. Int. Conf. Learn. Representations*, San Diego, CA, USA, 2015, pp. 1–14.
[63] G. Bertasius, L. Torresani, S. X. Yu, and J. Shi, "Convolutional random walk networks for semantic image segmentation," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 358–366.
[64] P. Pinheiro and R. Collobert, "Recurrent convolutional neural networks for scene labeling," in *Proc. 31st Int. Conf. Mach. Learn.*, 2014, pp. 82–90.
[65] Q. Guo-Jun, "Hierarchically gated deep networks for semantic segmentation," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2267–2275.

Lazhar Khelifi received the bachelor's degree in computer science from the Faculty of Sciences, Gafsa University, Gafsa, Tunisia, in 2010, the master's degree in computer science from the Faculty of Mathematical, Physical and Natural Sciences, University of Tunis El Manar, Tunis, Tunisia, in 2012, and the Ph.D. degree in computer science from the University of Montreal, Montreal, QC, Canada. His current research interests include image segmentation, fusion, multi-objective optimization, and machine learning.

Max Mignotte received the D.E.A. degree in digital signal, image, and speech processing from the Grenoble Institute of Technology, Grenoble, France, in 1993, and the Ph.D. degree in electronics and computer engineering from the Université de Bretagne Occidentale, Brest, France, and the Digital Signal Laboratory, French Naval Academy, Brest, France, in 1998. He was an INRIA (French Institute for Research in Computer Science and Automation) Postdoctoral Fellow with the Département d'informatique et de recherche opérationnelle, University of Montreal, Montreal, QC, Canada, from 1998 to 1999. He is currently an Associate Professor with the Computer Vision and Geometric Modeling Laboratory, University of Montreal. His current research interests include statistical methods, Bayesian inference, and hierarchical models for high-dimensional inverse problems, such as segmentation, parameters estimation, fusion, shape recognition, deconvolution, 3-D reconstruction, and restoration problems. He is also a member of the Laboratoire de Recherche en Imagerie et Orthopédie, Center de Recherche du Center Hosp italier de l'Université de Montréal, Hôpital Notre-Dame, Montreal, and a Researcher with CHUM.