16th European Signal Processing Conference (EUSIPCO 2008), Lausanne, Switzerland, August 25-29, 2008, copyright by EURASIP

UNSUPERVISED HIERARCHICAL IMAGE SEGMENTATION BASED ON THE TS-MRF MODEL AND FAST MEAN-SHIFT CLUSTERING

Raffaele Gaetano[†]*, Giuseppe Scarpa*[†]*, Giovanni Poggi*[†]*, and Josiane Zerubia*[‡]

[†]Dip. Ing. Elettronica e Telecomunicazioni,
Università "Federico II" of Naples, Italy
phone/fax: + (39) 0817683151/49,
email: *firstname.lastname@unina.it*

[‡]Ariana Research Group,
INRIA - I3S, Sophia Antipolis, France
phone/fax: + (33) 492387865/7643,
email: *josiane.zerubia@sophia.inria.fr*

ABSTRACT

Tree-Structured Markov Random Field (TS-MRF) models have been recently proposed to provide a hierarchical multiscale description of images. Based on such a model, the unsupervised image segmentation is carried out by means of a sequence of nested class splits, where each class is modeled as a local binary MRF.

We propose here a new TS-MRF unsupervised segmentation technique which improves upon the original algorithm by selecting a better tree structure and eliminating spurious classes. Such results are obtained by using the Mean-Shift procedure to estimate the number of pdf modes at each node (thus allowing for a non-binary tree), and to obtain a more reliable initial clustering for subsequent MRF optimization. To this end, we devise a new reliable and fast clustering algorithm based on the Mean-Shift technique. Experimental results prove the potential of the proposed method.

1. INTRODUCTION

Along with the advances of research in the image analysis and processing fields, the problem of segmentation is assuming an ever growing importance in many applications, such as medical image analysis, remote-sensing image classification, content based image retrieval, etc. Given the large-spectrum goal of image segmentation, that is, providing a partition of image pixels into some regions according to certain homogeneity criteria, it is easily understood that such a problem can be addressed with a wide variety of approaches. This leads to application-specific solutions that can also make sense at different levels of abstraction. In this widely varying scenario, MRF-based image modeling, first introduced in the 80's [1], still remains a very popular approach, mainly because of its effectiveness and flexibility in defining local dependencies among adjacent pixels, thus encompassing prior knowledge in the segmentation process with a reasonable complexity.

In order to improve the description capabilities of conventional MRF models and reduce the overall complexity of the derived segmentation algorithms, a new hierarchical MRF has been recently proposed, the Tree-Structured MRF (TS-MRF) model [2, 3], that proved quite effective and reliable, especially for the analysis and classification of remote sensing images. Such a model is motivated by the observation that images are often characterized by a distinctive hierarchical structure, with regions that interact with one another in different ways and at different scales of observation. The TS-MRF allows to model such a behavior by defining a suitable tree structure for the image of interest, and associating with each inner node of the tree a different image region and a different MRF, which is completely local to the corresponding region and has its dedicated parameters. This approach guarantees a much higher local adaptivity than classical MRFs. In addition, the segmentation problem can be formulated recursively, reducing a general K-ary segmentation procedure to a sequence of steps with just a few classes each, with a significant reduction of complexity.

Segmentation based on the TS-MRF model has proven very successful in the supervised case [3], when the number of classes of interest and their synthetic parameters are known *a priori*. In the unsupervised case [2] results are also good, especially if compared with those of unstructured techniques, but some critical issues remain to be addressed. In fact, lacking any prior information, one is forced to estimate, by recursive optimization at each node, the very same tree structure underlying the data. If the optimization is inaccurate at some nodes, the whole tree structure might deviate from the most suitable one, with various undesirable effects, like fusion of different classes or oversplitting of others.

In this work we propose an improved version of the TS-MRF unsupervised segmentation algorithm that addresses the major problems briefly outlined above. The main improvements come from the use of a Mean-Shift based clustering. As a matter of fact, the Mean-Shift procedure [4] was already used in [5] to detect the number of modes, and hence the number of children for each node of the tree. Here, however, its use is carried further, and besides finding the dominant modes for each class, it replaces the *Generalized Lloyd Algorithm* (GLA) [6] as the initial clustering technique, providing a much more reliable starting point for the subsequent MRF-based segmentation, and a much easier and stable detection of the correct tree-structure for the data. This is obtained through some significant modification of the Mean-Shift clustering itself, which now makes use of a variable-bandwidth strategy based on the *k-Nearest Neighbour* (k-NN) technique, and is implemented with a speed-up strategy that cuts significantly the computational complexity, otherwise intolerable for such applications.

In section 2, we first recall the basics of Mean-Shift analysis, and then describe the new Mean-Shift clustering algorithm, focusing in turn on the variable-bandwidth strategy, and on the speed-up solutions introduced. In Section 3, after describing in more details the TS-MRF model, and the related segmentation algorithm, we show how the new clustering tool can be used to improve the performance for unsupervised segmentation tasks. Finally, Section 4 provides experimental evidence of the improved performances and draws conclusions.

2. FAST VARIABLE-BANDWIDTH MEAN-SHIFT CLUSTERING

2.1 Background

The Mean-Shift procedure for mode detection [4] is a robust and effective tool to compute local maxima of a probability distribution over a given feature space, based on the well known *Parzen Window* framework [7] for non-parametric density estimation.

The rationale behind this algorithm is that samples in a certain feature space can be easily associated with an empirical probability density function: briefly, if we consider a d-dimensional feature spaces and a set of n data points $\{s_i\}_{i=i}^{n}$, the following expression can be a reasonable estimation for the pdf:

$$\hat{p}(s) = \frac{c_{K_p}}{nh^d} \sum_{i=1}^{n} K_p \left(\left\| \frac{s - s_i}{h} \right\|^2 \right), \tag{1}$$

where $K_p(\cdot)$ is a univariate strictly positive kernel profile function, such that a radially symmetric kernel can be generated from it through a rotation in $\mathscr{R}^d$, c_{K_p} is a normalizing constant and h is the kernel size, often indicated in the literature as the "bandwidth" parameter, that controls the resolution at which modes are detected. It is demonstrated that the gradient of this expression can be written as $\nabla \hat{p}(s) = q(s)\mathbf{m}_{h,g}(s)$, where $q(\cdot)$ is a scalar function and

$$\mathbf{m}_{h,g}(s) = \frac{\sum_{i=1}^{n} s_i g\left(\left\| \frac{s-s_i}{h} \right\|^2 \right)}{\sum_{i=1}^{n} g\left(\left\| \frac{s-s_i}{h} \right\|^2 \right)} - s \tag{2}$$

is the so called *mean shift* vector, where $g(\cdot) = -K_p'(\cdot)$. Therefore, based on the fact that mean shifts always point towards the maximum increase in the density, a gradient ascent procedure can be run, starting from any data point of the sample set, that will eventually converge to a stationary point in the distribution, that is, a mode of the pdf. Once a starting kernel center s is assigned, the procedure consists of two iterative steps:

1. compute the mean shift vector $\mathbf{m}_{h,g}(s)$,
2. update the kernel center $s = s + \mathbf{m}_{h,g}(s)$,

Clearly, to detect all significant modes, this procedure must be executed many times, each time with a different initialization, in order to cover most of the feature space.

2.2 Clustering by the mean-shift

The detection of modes through the Mean-Shift procedure determines an implicit clustering strategy over the feature space, since all the starting points of trajectories that converge towards the same mode (that is, belong to its "basin of attraction") form a well defined cluster. However, this would require running the Mean-Shift procedure for each point of the feature space, so as to identify the basin of attraction of all modes as clusters. Of course, this is unfeasible in practice, since for sample sets larger than several hundreds of data points computational time becomes extremely large for most of the possible applications. Hence, an efficient implementation is usually required, especially for data-intensive cases.

Another critical implementation issue is the choice of the kernel size, or *bandwidth* parameter, which plays a central role for density estimation since it determines the smoothness of the pdf and, consequently, the number of modes that the algorithm singles out[1]. Using a too large bandwidth leads to underestimating the number of modes, and the opposite for too small a value.

We propose here an implementation of Mean-Shift clustering which addresses the two problems briefly outlined above. In particular, the new algorithm is based on

- a data-dependent adaptive kernel size h that overcomes the instability observed for example in [5];
- a fast clustering technique that enables its use for real-world applications.

2.2.1 K-NN *based adaptive bandwidth selection*

The original Mean-Shift procedure proposed by Comaniciu [4] uses a fixed bandwidth parameter h, but this is clearly inappropriate when the density of points in the feature space varies wildly. In such cases, in fact, no value can be well suited for both high- and low-density areas.

To face this problem, we adapt the bandwidth parameter locally in the feature space by taking into account only the first k-Nearest Neighbors in the computation of the Mean Shift vector. This amounts to truncating the kernel at some distance from the center but, if k is not too small, this truncation will take place when the kernel has already a negligible value, independent of the local density. The bandwidth, instead, will clearly depend on the local density, being larger in low-density areas and smaller in high density ones.

In more detail, given a suitable value of k, at each step of the procedure the set $NN(s)$ of k points closest to s is singled out, and the kernel size is calculated as

$$h(s) = \sqrt{\frac{1}{k} \sum_{i \in NN(s)} \|s - s_i\|^2}, \tag{3}$$

This value is then used in (2) for the computation of the mean-shift vector where the summation is again restricted to the points in $NN(s)$.

It could be observed that this solution moves the problem from the estimation of parameter h, to that of parameter k, but it is well-known [8] that *k-NN* estimation is quite robust w.r.t. its parameter, and works quite well also in spaces of high dimensionality, which are instead quite challenging for the Mean-Shift. In next section, we propose a data dependent procedure for obtaining a stable estimate of the k parameter.

2.2.2 *Fast clustering strategy*

Our speed-up strategy is based on the obvious consideration that all points that lie on the trajectory that goes from the starting point to the corresponding mode belong necessarily to the same basin of attraction. Therefore, they could all be attributed, without error, to the same cluster.

Although it is extremely difficult that any sample point will coincide *exactly* with a point of this path, one can reasonably assume that sample points that are *close* to the trajectory belong very likely to the same basin. By clustering all such points at once we drastically reduce the complexity, but also risk to cause some errors, especially for data points that are close to the watershed between two basins of attraction. Hence, in order to preserve the accuracy of clustering,

[1] Less critical is the kernel function, for which we select the quite usual Gaussian shape.

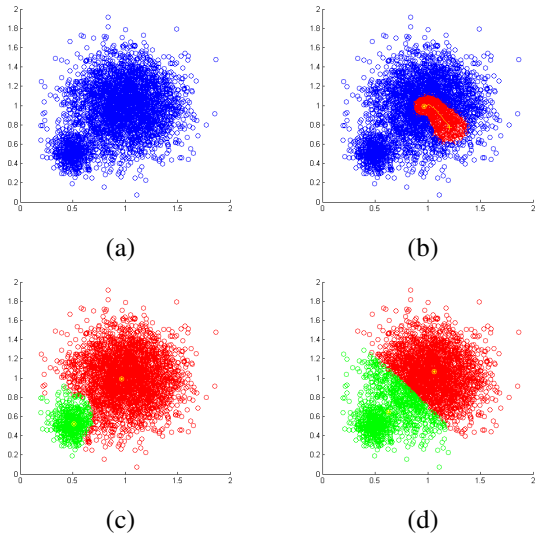


Figure 1: (a) bi-modal sample set, (b) Mean Shift trajectory with the corresponding "voting" points, (c) final clustering, (d) GLA-based clustering for comparison.

we do not assign sample points on the fly, but rather implement a voting mechanism and decide only a posteriori, with a majority rule, when all sample points have been touched by at least one trajectory. The modified procedure can be summarized as follows:

1. *Initialization*: set all sample points as *non visited*.
2. *Mean-Shift*: run the procedure starting from a randomly selected non visited point: at each step along the trajectory, mark as *visited* all points s_i such that $\|s-s_i\| < h(s)$, and for each of them add a vote for the "final" mode.
3. *Mode validation*: once convergence is reached, compute the distance $d_{\min}$ between the new tentative mode and the closest mode already detected:
   - if $d_{\min} < h/2$ reject the new mode, and mark the closest mode as final;
   - otherwise accept the new mode, and mark it as final.
4. *Test*: if there are still non visited points, go to step 2.
5. *Clustering*: assign each visited point to the mode (and cluster) with the most votes.

An example of clustering provided by the described procedure is presented in Fig.1: the bivariate sample set of part (a), obtained as a mixture of two normally distributed data sets, is given as input to the clustering algorithm. In part (b) the effect of a single modified Mean-Shift procedure is represented, where all the points in red are "giving a vote" to the final mode. Part (c) shows the final clustering, which appears to follow quite faithfully the underlying distribution and is certainly much better than the GLA-based clustering shown in part (d) where, in addition, the correct number of clusters had to be provided as a further input.

3. TS-MRF UNSUPERVISED SEGMENTATION BASED ON MEAN-SHIFT CLUSTERING

As already mentioned in Section 1 and discussed in detail in [2, 3], TS-MRF modeling is based on the hypothesis that the data possess an inherent hierarchical structure in terms of spatial and spectral properties. Given *a priori* the number of classes K to be retrieved with their parameters, and a suitable tree T that describes the hierarchical structure of data, a "simple" MRF is associated with each inner node t and the segmentation can be carried out by top-down induction over the tree with a recursive optimization[2] of the different MRFs.

In the unsupervised case, however, no prior information is available on the image, and all parameters, including the tree structure, must be retrieved during the process. the segmentation is decomposed into a sequence of nested splits, starting from the whole image, and going on until all elementary regions are identified according to a given stopping criterion. The entire image is therefore associated with the root of a tree, and each split creates some new nodes, generating gradually the desired tree structure whose growth is governed by a suitable metric called *split-gain* [2]. Terminal nodes of the structure correspond to the final classes of the map. The final product is a hierarchical multiscale image, with a synthetic high-level description provided by the tree itself together with the class parameters, and a *set* of finer and finer maps left to the user's selection.

Of course, a number of implementation compromises, often driven by complexity concerns, impact on the overall performance. One such choice is to consider only binary tree structures, reducing the segmentation process to a sequence of nested binary splits controlled by a suitable stopping criterion. Such a constraint, however, might cause the detection of false contours as can happen when three or more balanced classes are present in the same region. In [5] we removed this constraint by resorting to the Mean-Shift procedure to detect the number of pdf modes in a class, and hence the number of children at a node. Another questionable choice is the use of the GLA to carry out the initial clustering needed to perform the MRF optimization at each node. In fact, image pixels are often described by a complex and generally unbalanced probability distribution in the spectral domain, in which case the GLA can easily provide inaccurate results, as in the example of Fig.1(d).

Therefore, we now replace the GLA based clustering with the more accurate variable-bandwidth Mean-Shift based clustering described in the preceding section. Even though our fast implementation helps limiting the processing weight, plain Mean-Shift clustering would have an exceedingly high computational complexity for the very large images we usually deal with, and hence we will eventually resort to a hybrid *Mean-Shift/Maximum Likelihood* (MS-ML) classifier. In more details, for each region to split, we extract a reasonably large random subset of pixels (from 1% of the region area to the entire region, depending on its size); the Mean-Shift clustering described in Section 2 is then applied to this sample set, to retrieve the number of classes and their statistics; this information is eventually used by a Maximum Likelihood classifier that runs over the whole region.

In Section 2, we did not address the problem of selecting a suitable value of k for the *k-NN* based bandwidth estimation of (3). A typical choice is to set k to a fraction, *e.g.*, 10%, of the sample set cardinality. Although this rule works usually well, thanks to the robustness of *k-NN*, unreliable results are sometimes obtained, such as an unlikely proliferation of modes. This is not surprising, given that the same algorithms are used for all nodes, from the root, correspond-

[2] In this work we refer to the Potts MRF framework, and optimization is obtained using a *Maximum a Posteriori* criterion, see [9, 10].

(a) (b) (c)

Figure 2: Detail of the XS3 channel (©SPOTImage/CNES) (a), initial *sea class* split using GLA (b), and MS-ML (c).

ing to the whole image, to terminal leaves corresponding to much smaller and more fragmented regions.

Therefore, we use a simple heuristic procedure that adapts the value of k to minimize such unlikely behaviors. Our underlying assumption is that, most of the times, the data structure can be well described through one or more binary splits. Therefore, we set the initial value of k as $k_0 = \text{round}(\alpha_0|S|)$, with S the selected sample set. Then the Mean-Shift procedure, *without indirect clustering*, is run for a maximum of N_1 times, keeping track of the number D of detected modes: if D remains stable at 2 for N_2 times, then the current value of k is accepted, while if $D > 2$ [$D < 2$], then k is increased [decreased] by the quantity $\Delta \cdot k$ and the procedure is repeated. In any case, k is not allowed to escape a certain range $[\alpha_1 k_0, \alpha_2 k_0]$, when it is freezed anyway. This procedure provides a solid criterion to decide whether to split a node or not, since the stable detection of a single mode qualifies the corresponding region as elementary.

Using a more reliable technique to carry out the initial clustering does certainly improve the subsequent MRF optimization, but there is a more subtle and important consequence in the context of hierarchical segmentation. In fact, the MS-ML clustering provides quite a reliable segmentation in the spectral domain, while the MRF model allows to take into account contextual information to regularize the final map. The points that change label during MRF optimization turn out to be "outliers" in the spectral domain for the final class ω, that is, their statistics will be far apart from those of points originally attributed to ω by the MS-ML technique. If class ω is segmented again, such outliers can give origin to one or more separate clusters, leading to dramatic oversegmentation errors. We are now in the position to avoid this dangerous phenomenon, by simply erasing such points from the new sample set. Notice that this was not possible with a GLA initialization, since the initial clusters were so far from the final segmentation (see Fig.1) that such erasure would eliminate large chunks of valid data.

4. EXPERIMENTAL RESULTS

We assess the performance of the improved unsupervised segmentation algorithm through a set of experiments on a SPOT satellite image of the Lannion Bay, in France, August 1997 (©SPOTImage/CNES), composed of three 1480×1024 bands and a spatial resolution of 20m.

For both the original TS-MRF algorithm and the new version proposed here we use the same settings for the MRF optimization part, and stop the tree growth at 8 classes. The mode detection procedure uses $\alpha_1 = 0.08, \alpha_2 = 0.12, \Delta = 0.05, N_1 = 20$ and $N_2 = 10$.

The improvements due to the use of the MS-ML are quite clear since the first stages of segmentation. In Fig.2(a) we show a detail of the source image, along with two maps that, for both the original (b) and new version (c) of the algorithm, show the "sea" class (in white) as identified by the top-level clustering, before any MRF regularization. The errors introduced by the GLA are quite evident in Fig.2(b), as well as the very high accuracy of the MS-ML classification of Fig.2(c). Such a good initialization will likely improve, and certainly simplify the subsequent optimization process (making up for the increased complexity of the MS-ML clustering). Moreover, it will allow to single out easily the few label-switching points to eliminate in further spectral clustering steps.

	TS-MRF w/GLA		**TS-MRF w/MS-ML**	
Classes	*U.A.*	*P.A.*	*U.A.*	*P.A.*
Water	100.0%	75.6%	100.0%	94.2%
Bare Soil	75.5%	95.0%	97.5%	69.1%
Urban	4.0%	6.3%	49.2%	40.8%
Forests	97.1%	46.2%	97.1%	95.6%
Temp. Mead.	38.5%	30.2%	25.3%	33.2%
Perm. Mead.	23.1%	22.0%	33.7%	28.8%
Vegetables	0.0%	0.0%	3.4%	4.0%
Corn	63.9%	95.5%	65.2%	94.5%
Overall Acc.	**59.8**%		**74.4**%	

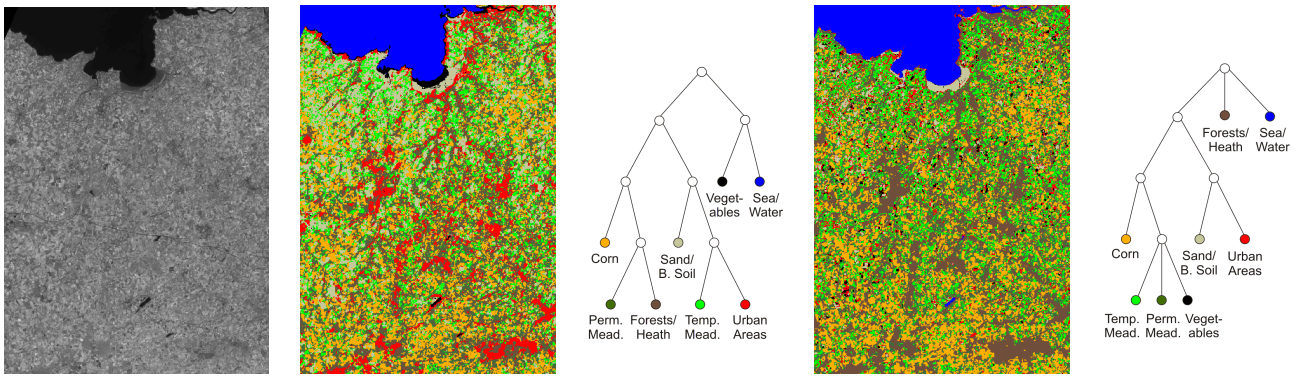
Table 1: Per-class and overall accuracies for the classification of Fig.3(b) and Fig.3(c) respectively.

Fig.3(a) shows the complete image (again XS3 channel), provided with the available ground truth (©COSTEL), not reported for sake of brevity, used to compute quality figures. The segmentation maps obtained with the original and improved TS-MRF algorithms are reported in Fig.3(b) and (d), respectively. Fig.3(c) and (e) instead, show the tree structures detected by both algorithms, where the leaves are associated *a posteriori* to the eight semantic classes so as to maximize the overall accuracy as computed on the ground truth.

At a visual analysis, results provided by the proposed version are much more accurate than those of the original algorithm: no major losses are noticeable, at least on top level classes, unlike in the map of Fig.3(c) where a serious oversplitting of the "forests" class sticks out. Numerical results confirm such empirical observations: the overall classification rate goes from around 60% to 74.4% mainly due to the more precise detection of some large classes, such as the "forests" and "urban areas" classes, as appears from the user's and producer's accuracies[3] reported in Tab.1. Such an improvement can be likely ascribed to the improved segmentation accuracy obtained in the first steps, also due to the more flexible tree structure. As can be seen in Fig.3(e), in fact, the new technique, by resorting directly to a 3-class top-level split, immediately detects and validates the "forests" class, preventing it from being oversplit in later stages.

Another interesting result concerns the TS-MRF *supervised* segmentation technique, referred as TS/U in [3]. The supervised procedure has been run here replacing the original binary tree-structure selected in [3] by visual inspection with the tree structure of Fig.3(e) detected by the unsupervised technique proposed in this paper. Quite an accurate segmentation map has been obtained, with an overall accuracy of 85.3% as opposed to the 82.3% obtained with the

[3]Complete confusion matrices could not be shown for lack of space.



(a) (b) (c) (d) (e)

Figure 3: XS3 channel of an image of Lannion Bay, France (©SPOTImage/CNES) (a), unsupervised segmentation by the original TS-MRF algorithm (b) and the corresponding tree (c), the new version (d) and the corresponding tree (e).

Figure 4: Segmentation maps obtained on two images from the Berkeley dataset (top row), using the new version of the TS-MRF algorithm (middle) and the original one (bottom).

hand-picked tree-structure. Therefore, the tree-structure detected here does fit well the source data and can be used as a preliminary tool in supervised TS-MRF segmentation, eliminating the need for a heavy user intervention.

Finally, in Fig.4 we show some segmentation maps obtained for two images (top row) of the well-known Berkeley dataset [11], widely used for benchmarking object-oriented techniques. Although the TS-MRF model is oriented to recognizing "classes" rather than objects, and should be suitably modified to work with this type of applications, we point out the improvements granted by the new version (middle) over the original one (bottom row), such as the more accurate recognition of the roof (left image), completely lost using the original technique, or the snake (right), originally assigned to the same color class with part of the background.

Such results, though only partial, indicate that the new TS-MRF segmentation algorithm, with Mean-Shift clustering for mode detection and MRF initializations, guarantees in general a more stable and accurate segmentation.

REFERENCES

[1] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 1, pp. 721–741, Nov. 1984.

[2] C. D'Elia, G. Poggi, and G. Scarpa, "A tree-structured Markov random field model for Bayesian image segmentation," *IEEE Transactions on Image Processing*, vol. 12, no. 10, pp. 1259–1273, October 2003.

[3] G. Poggi, G. Scarpa, and J. Zerubia, "Supervised segmentation of remote sensing images based on a tree-structured MRF model," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 43, no. 8, pp. 1901–1911, August 2005.

[4] D. Comaniciu and P. Meer, "Mean Shift: a robust approach toward feature space analysis," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 24, no. 5, pp. 603–619, May 2002.

[5] R. Gaetano, G. Poggi, and G. Scarpa, "Identification of image structure by the Mean Shift procedure for hierarchical MRF-based image segmentation," in *Proc. EUSIPCO 2006*, Florence, Italy, Sept. 2006.

[6] A. Gersho and R. Gray, *Vector quantization and signal compression*. Boston, MA: Kluwer, 1992.

[7] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*. Wiley, 2000.

[8] D. W. Scott, *Multivariate Density Estimation*. Wiley, 1992.

[9] G. Winkler, *Image analysis, random fields and dynamic Monte Carlo methods*, 1st ed. Springer-Verlag, 1995.

[10] S. Z. Li, *Markov random field modeling in image analysis*, 1st ed. Springer-Verlag, 1995.

[11] 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'l Conf. Computer Vision*, vol. 2, July 2001, pp. 416–423.