PERFORMANCE EVALUATION OF PROBABILITY DENSITY ESTIMATORS FOR UNSUPERVISED INFORMATION THEORETICAL REGION MERGING

Felipe Calderero,   Ferran Marques

Department of Signal Theory and Communications
Technical University of Catalonia (UPC)
Barcelona, Spain

Antonio Ortega

Department of Electrical Engineering
University of Southern California
Los Angeles, California, USA

ABSTRACT

Information theoretical region merging techniques have been shown to provide a state-of-the-art unified solution for natural and texture image segmentation. Here, we study how the segmentation results can be further improved by a more accurate estimation of the statistical model characterizing the regions. Concretely, we explore four density estimators that can be used for pdf or joint pdf estimation. The first three are based on different quantization strategies: a general uniform quantization, an MDL-based uniform quantization, and a data-dependent partitioning and estimation. The fourth strategy is based on a computationally efficient kernel-based estimator (averaged shifted histogram). Finally, all estimators are objectively evaluated using a database with available ground truth partitions.

Index Terms— Density estimation, statistical models, region merging, image segmentation.

1. INTRODUCTION

Image segmentation can be considered as a first and key step into image analysis and pattern recognition. Commonly, a unique solution for the image segmentation problem does not exist. To overcome this situation, a hierarchical segmentation approach can be used where, instead of a single partition, a hierarchy of partitions is provided. An important type of hierarchical segmentation approaches are region merging techniques. Starting from an initial partition, regions are iteratively merged until a stopping criterion is reached. Thus, region merging algorithms are specified by: a merging criterion, defining the cost of merging two regions; a merging order, determining the sequence in which regions are merged based on the merging criterion; and a region model that determines how to characterize a region.

Unsupervised information theoretical region merging (IT-RM) techniques were presented in [1, 2] as a unified solution to natural and texture image segmentation. Formally, IT-RM techniques use a statistical region model, leading under certain principles (such as maximum likelihood and minimum classification error) to information theoretic statistical measures as merging criteria, which, in turn, can be combined with a maximum similarity (classical) or a scale-based merging order. In our previous work, merging criteria and merging order were objectively and subjectively evaluated in terms of segmentation error using as statistical model estimator a uniformly quantized version of the region empirical distribution.

These methods require models of pixel data for which, in general, regions do not provide sufficient samples. This density estimation problem is the fundamental issue we are dealing with in

this work. First, we provide evidence that using directly the (non-quantized) empirical distribution of the region does not lead to the best segmentation results. Second, we study the variation of the performance for four different probability density and joint probability density estimation techniques. We compare adaptive quantization-based to kernel-based approaches, concluding that the smoother and more detailed solution provided by the kernel-based estimators is clearly superior. Results are relevant for other image processing techniques using non-parametric density function estimation (as mean-shift object tracking [3] or content-based image retrieval [4]).

The paper is structured as follows. Section 2 briefly reviews the IT-RM techniques. Section 3 outlines the main problems of the direct use of the original region distributions. In Section 4 two quantization-based approaches are presented: a uniform quantization with a variable number of bins for each image based on the minimum description length principle [5]; and a non-uniform image quantization strategy based on a data-dependent partitioning of the empirical distribution [6]. Section 5 presents a computationally efficient approximation of continuous kernel-based estimators [7]. An objective comparison of all estimators is performed in Section 6. Finally, conclusions are outlined in Section 7.

2. INFORMATION THEORETICAL REGION MERGING

IT-RM techniques (outlined in Figure 1) have been developed under a statistical framework, leading to (i) a statistical region model, (ii) merging criteria based on similarity measures between probability distributions, and (iii) merging orders based either exclusively on the maximum similarity or combined with scale information.

Depending on the statistical assumptions, two region models have been proposed. First, when the region pixels are considered independent and identically distributed (i.i.d.), the region model is the discrete probability density function for its pixel values [1]. Second, when region pixels are considered statistically dependent, the region model is the probability transition matrix of the Markov process [2].

Merging criteria have been developed under two different principles. A merging criterion based on the Kullback-Leibler (KL) divergence is obtained when we merge at each step the pair of regions with maximum likelihood to have been generated by the same statistical model. In turn, when merging the pair of regions with maximum classification error, the resulting criterion is based on the Bhattacharyya (BHAT) coefficient.

Finally, two merging orders have been proposed. A classical maximum similarity order, that is, merging the pair of regions with maximum similarity in terms of the specified merging criterion (*area-weighted*); and a scale-based order, where the scale information is used to merge with highest priority regions with areas too small to be significant at each analysis level (*area-unweighted*).
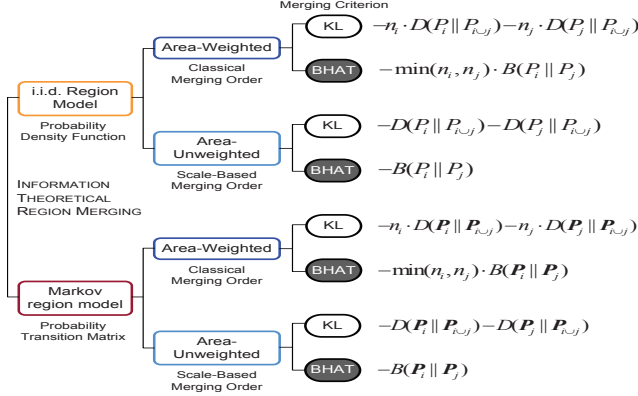
Fig. 1: Information theoretical region merging techniques.

The combination of the proposed region models, merging criteria and merging orders leads to eight different techniques. They are summarized in Figure 1, where n, P, and $\mathbf{P}$ refer to the area, the probability density function, and the probability transition matrix of a region, respectively. The Kullback-Leibler divergence and the Bhattacharyya coefficient are represented as $D(\cdot||\cdot)$ and $B(\cdot||\cdot)$.

3. DENSITY ESTIMATION PROBLEMS WITH A REDUCED NUMBER OF SAMPLES

In our previous work, the statistical region models were approximated by the empirical distribution (normalized histogram) and the empirical probability transition matrix (cooccurrence matrix) of the pixel data. Although one might think that using non-quantized empirical distributions would provide better segmentation results, we observed that quantization can reduce both under- and oversegmentation errors (see Section 6). A similar conclusion was reported in [8] for cooccurrence matrices in the context of texture analysis.

This fact is a consequence of the limited number of pixels to estimate the region statistical models. This leads to the *zero probability problem* in the estimated densities (bins with zero probability) which, in practice, may decrease the value of the similarity measure between the densities (as illustrated in Figure 2a). In addition, it may lead to a lack of overlapping between similar distributions. As the statistical measures used as merging criterion only consider the overlapping probabilities between the distributions, but not the distance (in terms of bins) between them, if two contiguous estimated distributions belonging to a single real density do not overlap, their similarity is exactly the same as two strongly separated distributions (see Figure 2b).

In [1, 2], a uniform quantization approach with a constant number of bins was introduced to palliate these problems. Intuitively, a moderate degree of quantization will reduce the number of bins with zero probability and may help to overlap the distributions and facilitate their merging. This concept is evaluated in Section 6.

Apart from the uniform quantization strategy using a fixed number of bins, we explore the performance of more sophisticated density estimators. These approaches are described in the next sections.

4. ADAPTIVE QUANTIZATION-BASED APPROACHES

4.1. MDL-Based Uniform Quantization
A uniform quantization of the original empirical pixel distribution is used, but instead of fixing the number of bins, the degree of quantization is determined by the characteristics of the image. Concretely, we apply the minimum description length (MDL) principle to determine an appropriate number of bins, through the minimization of the stochastic complexity of the image, as proposed in [5].
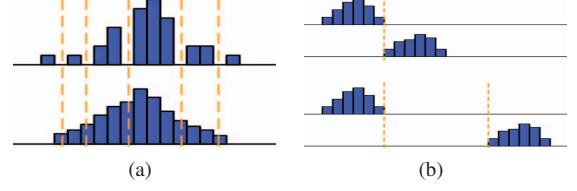

Fig. 2: Density estimation problems. (a) Zero probability problem in the empirical distribution. (b) Non overlapping close distributions.

As proved in [5], the MDL principle recommends choosing the number of bins M^* as to maximize the following expression:

$$M^* = \arg \max_M \sum_{\forall i} [(n_i - \tfrac{1}{2}) \log(n_i - \tfrac{1}{2})$$
$$- (N - \tfrac{M}{2}) \log(N - \tfrac{M}{2})) + N \log(M) - \tfrac{M}{2} \log(N)] \quad (1)$$

where n_i is the number of samples which take values in ith bin, N is the total number of samples, and M is the total number of bins.

For color (or multichannel) images, the MDL-optimal number of bins for each channel is determined, and then, the median value is selected to quantize all channels. For the Markov-based IT-RM techniques, the MDL-optimal number of bins is used for each dimension of the estimated cooccurrence matrix.

4.2. Data-Dependent Estimation
A universal estimator based on data space subdivision was proposed in [6] for estimating the Kullback-Leibler divergence $D(P||Q)$ between two arbitrary continuous distributions P and Q. The algorithm takes Q as the reference distribution and first divides the space into a given number of intervals such that each interval contains an equal number of samples of Q. Additionally, finer subdivisions may be performed to intervals where the density of P is high relative to the density of Q. The divergence between P and Q is estimated by the ratio between the empirical probabilities in each interval.

Besides its fast convergence to the real value of the divergence, it has been reported to perform well in applications where only a reduced number of samples is available [4]. The reason is that this approach assures that none of the resulting bins remains with zero probability. The generalization to multidimensional distributions (e.g., Markov-based IT-RM techniques) is straightforward [6].

In this work, two simplified versions of the data-dependent estimator proposed in [6] have been implemented. In both cases, for computational reasons, a single division (not further refinement) is applied. The algorithms are as follows (version name as in [6]):

(i) **Version A**: divides the space into M bins, such that each bin has l_M samples from Q, except possibly the last one.

(ii) **Version C**: divides the space into M bins, such that each bin has l_M samples from Q. Then, if an interval contains more than αl_M samples from P, it is further divided into intervals having αl_M, except possibly one of them.

As suggested in [6], the parameters are set to $l_M = \lfloor (N_Q)^\lambda \rfloor$ and $\alpha l_M = \lfloor (N_P)^\lambda \rfloor$, with $\lambda = 1/2$, where N_P and N_Q are the total number of samples from distribution P and Q, respectively. In addition, we explore other parameter setting, such as $\lambda = 1/4$, and $\lambda = 3/4$. For completeness, the same data-dependent division strategy is used for the techniques based on the Bhattacharyya coefficient, though no theoretical justification has been reported for the convergence of the estimator to the real value of this measure.

5. AVERAGED SHIFTED HISTOGRAM ESTIMATION

Kernel density estimators are non-parametric methods for estimating the probability density function of a random variable, generated as a linear combination of a continuous kernel function, K(x). Since

the complexity of a continuous Kernel estimator would be excessive in a region merging (iterative) process, we propose to use averaged shifted histograms (ASH), which can be seen as a computationally constrained version of kernel estimators. They combine a set of m histograms generated with a certain shift into the bin borders [7]. It can be shown that when the number of histograms tends to infinity, the ASH approximates the kernel estimator. In practice, ASH is implemented by generating a histogram with smaller bin width, and computing the discrete convolution with a triangular window:

$$f_{ASH}(k) = \frac{1}{nh} \sum_{i=1-m}^{m-1} \left(1 - \frac{|i|}{m}\right) p(k+i), \qquad (2)$$

where h and n are the original bin width of the averaged histograms and the total number of samples, respectively. Thus, the ASH estimator provides an averaged or low-pass filtered version of the histogram with fewer zero probability bins in the estimated distribution, and a larger tail (that may increase the overlapping between contiguous distributions). The ASH estimator can be generalized to the multidimensional case, using a multidimensional triangular window.

6. EVALUATION RESULTS

In all experiments in this section, IT-RM techniques[1] were applied on initial partitions with 1500 regions to ensure that all initial regions are large enough to initialize the statistical models. Partitions were computed using a step-wise optimal (size-weighted $\mathcal{L}_2$-norm) color-based region merging in YCbCr color space. The evaluation was performed over a subset of 100 images from the Corel$^{\copyright}$ image database [9]. Ground truth partitions were manually segmented in the context of the SCHEMA project (http://www.iti.gr/SCHEMA/).

To evaluate the quality of the partitions obtained using the different estimators, we use the distances defined in [10] and an extension of the methodology in [9]. First, an asymmetric distance is proposed: $d_{\text{asym}}(\Pi, \Lambda) \neq d_{\text{asym}}(\Lambda, \Pi)$. When Π is the partition to evaluate and Λ its ground truth partition, the first ordering measures the percentage of undersegmented pixels (undersegmentation error: UE), and the second, the oversegmentation error (OE) in Π with respect to the ground truth partition. Second, a symmetric distance, $d_{\text{sym}}(\Pi, \Lambda)$, is defined. When one of the partitions is the ground truth, this measure quantifies the global quality of a partition, and its trade-off between under- and oversegmentation.

The first relevant conclusion extracted from the current evaluation is that the performance of area-unweighted methods is specially sensitive to the probability distribution estimation. In fact, we have observed that the main differences into the segmentation results are not given by the region model (i.i.d. or Markov) or the merging criterion (KL and BHAT measures perform similarly, although in general BHAT is slightly superior), but by the merging order. For that reason, we will refer to the behavior of area-weighted and area-unweighted techniques for each density estimator. As a brief summary, area-weighted (area-unweighted) techniques present a lower UE (OE). Nevertheless, area-unweighted techniques show a better trade-off between both types of error, obtaining, for some specific techniques, comparable UE results (see Figure 3). The complete evaluation including all methods and all estimators can be found at *http://gps-tsc.upc.es/imatge/_Felipe/icip09/*. Due to space limitations, in this Section only the i.i.d. Bhattacharyya techniques (area-weighted and area-unweighted versions) are shown for comparison purposes in terms of UE and OE.

Uniform Quantization Evaluation: In our previous work, the uniform quantization strategy (with 5 and 10 bins) was shown to

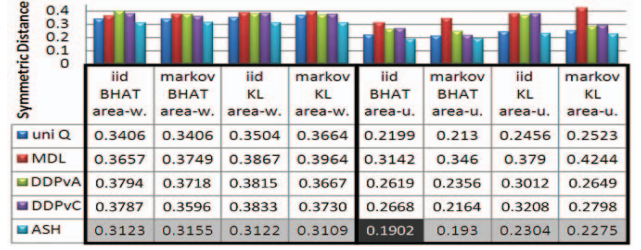|	iid BHAT area-w.	markov BHAT area-w.	iid KL area-w.	markov KL area-w.	iid BHAT area-u.	markov BHAT area-u.	iid KL area-u.	markov KL area-u.
uni Q	0.3406	0.3406	0.3504	0.3664	0.2199	0.213	0.2456	0.2523
MDL	0.3657	0.3749	0.3867	0.3964	0.3142	0.346	0.379	0.4244
DDPvA	0.3794	0.3718	0.3815	0.3667	0.2619	0.2356	0.3012	0.2649
DDPvC	0.3787	0.3596	0.3833	0.3730	0.2668	0.2164	0.3208	0.2798
ASH	0.3123	0.3155	0.3122	0.3109	0.1902	0.193	0.2304	0.2275

Fig. 3: Mean symmetric distance for the IT-RM techniques using fixed (uni Q) and MDL-adaptive (MDL) uniform quantization, version A and C of the data-depending quantization (DDPvA, DDPvC), and ASH estimator (ASH). For uni Q, ASH, and DDP versions, the parameter values with minimum error are considered.

outperform state-of-the-art segmentation techniques [1, 2]. Here, an evaluation in terms of UE, OE and global error as a function of the number of bins has been performed. Due to space limitations, UE and OE curves have not being included. They lead to the same conclusions exposed above. In Figure 3, the value of the symmetric distance minimum for all IT-RM techniques is shown.

For the sake of clarity, in the comparative figures shown for the other estimators, only the results corresponding to the optimal number of bins in terms of UE and of OE are shown.

MDL and Data-Dependent Evaluation: The second relevant conclusion of the evaluation is that the MDL-adaptive uniform quantization and the data-dependent estimation (Version A and C) perform similar, and with comparable results to the fixed uniform quantization approach. In the case of area-weighted methods, their performance is close to that provided by the optimal number of bins in terms of UE, but slightly worst in both UE and OE. For area-unweighted methods the differences are more significant (see Figure 4, 1st and 3rd rows for UE and OE, respectively). In some cases, MDL shows a better performance in terms of OE than the fixed uniform quantization. However, MDL and data-dependent strategies do not outperform in any case the uniform quantization approach in terms of global quality (see Figure 3). In addition, there is no significant variation of the data-dependent estimator performance when using the proposed λ values. For the sake of clarity, all results in this section correspond to the value of $\lambda=1/2$ suggested in [6].

ASH Evaluation: The third relevant conclusion is that ASH estimator outperforms the other estimation techniques. ASH estimator smooth densities (with a larger number of bins) are a better and more detailed approximation of the real region densities. This explains its superiority compared to the quantization-based techniques, that tend to piecewise constant density approximations.

Again, a larger performance variation is observed for the area-unweighted techniques than for area-weighted techniques when varying the m parameter (number of averaged histograms), as can be seen in Figure 4, 2nd (UE) and 4th (OE) rows. In this example, the averaged histograms correspond to the non-quantized empirical distributions of the pixel data (i.e. using 256 bins). The value of m is varied from 10 to 50 bins, which corresponds to averaging from 4% to 20% of the number of bins into the histogram. For area-weighted techniques, the ASH estimator provides a limited performance improvement, reducing around a 6% the UE and a 2% OE in mean, with respect to the optimal uniform quantization value. OE decreases as m increases, while UE shows a minimum around $m = 40$. Nevertheless, compared to the local minimum of the symmetric distance present around 5-15 bins for the uniform quantization case, the ASH estimator for $m = 40$ decreases the global partition error into a 10% in mean.
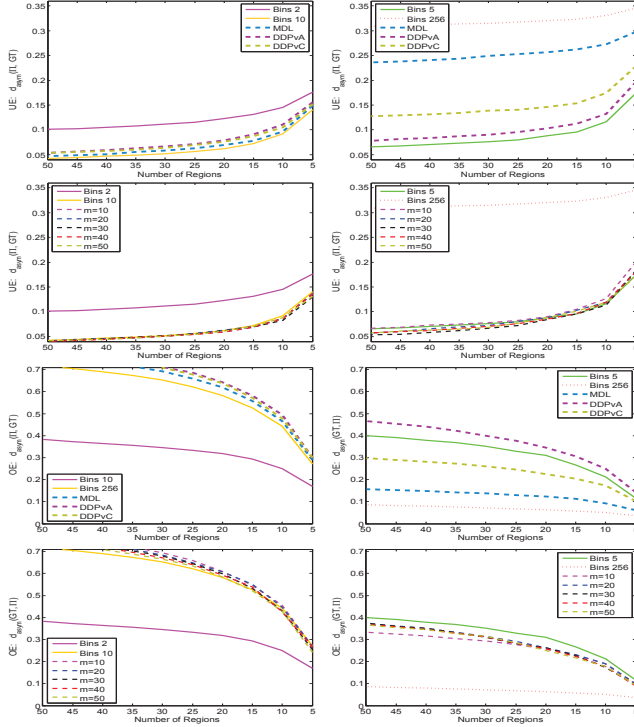
---

[1]For area-unweighted methods the scale parameter α was set to 0.07 [1].

Fig. 4: Performance comparison of density estimator for the i.i.d. Bhattacharyya area-weighted (1st column) and area-unweighted (2nd column) techniques. Mean UE and OE for: MDL uniform quantization and data-dependent version A (DDvA) and version C (DDvC) estimation (1st and 3rd rows); ASH estimator for different number of averaged histograms, m (2nd and 4th rows). Curves corresponding to the optimal number of bins in terms of UE and of OE for the fixed uniform quantization are also shown.

For area-unweighted methods the improvement is more significant, around 15% of the UE and 14% of the OE. The UE decreases as the number of averaged histograms (m) increases for values under $m = 30$. For larger values, 30-50, similar values of UE are obtained. OE behaves in the opposite way, increasing as m increases. This leads to the best trade-off in terms of global quality for m around 40, that improves in more than 9% the global quality of the partitions. A complete comparison in terms of symmetric distance with the other estimators is outlined in Figure 3 and fully illustrated at *http://gps-tsc.upc.es/imatge/_Felipe/icip09/*.

Figure 5 compares the partitions obtained by a uniform quantization approach and an ASH estimator, using the parameter values minimizing the symmetric distance in each case (5 bins and $m = 40$, respectively). ASH partitions show more regular and accurate contours than the uniform quantization partitions and less oversegmentation, as contours due to quantization effects do not appear into the ASH partitions (for instance, see the segmentation of the lake in the image in the 1st row, or the sky in the 3rd and 4th rows of Figure 5).

The increase of computational time of the ASH estimators with respect to a fixed uniform quantization can be controlled by m and by the number of bins in the original histogram. For instance, the IT-RM technique selected to compute the examples in this section is 5 times slower using the ASH estimator with $m = 40$ and histograms of 256 bins than using a uniform quantization with 5 bins, 2.3 times slower using 128 bins and $m = 20$, and less than 2 times for 64 bins and $m = 10$, without a significant variation of its performance.
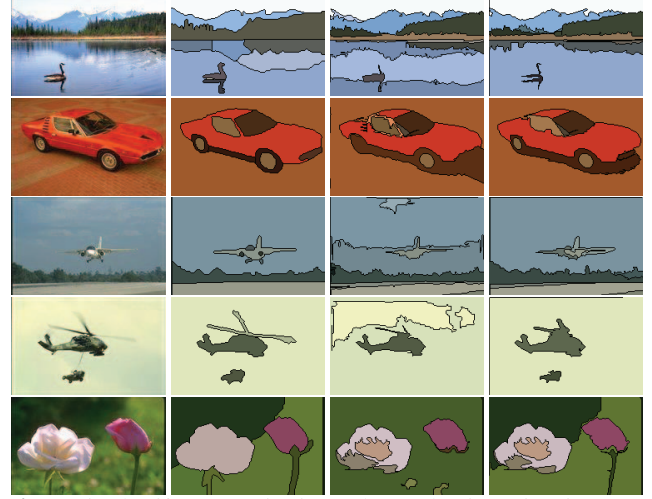

Fig. 5: Fixed uniform quantization and ASH density estimation comparison (i.i.d. Bhattacharyya area-unweighted technique). Columns from left to right: original image, human partition, partition computed using uniform quantization (5 bins), and using ASH estimator ($m = 40$), both with same number of regions as human partition.

7. CONCLUSIONS

The first conclusion is that the performance of area-unweighted methods is specially sensitive to the probability distribution estimation. The evaluation also shows that MDL uniform quantization and data-dependent estimators do not provide an advantageous approach since a computationally simpler strategy, such as a fixed uniform quantization, presents a similar or better performance. On the contrary, ASH estimators are a valuable solution in applications where the segmentation errors may be crucial or not being computationally constrained, since an increase in computational time (which can be controlled by m and the number of bins in the original histogram) provides a significant improvement in terms of undersegmentation, oversegmentation and global errors. Results are useful for other image processing techniques using non-parametric density estimation and facing similar problems due to the lack of available samples.

8. REFERENCES

[1] F. Calderero and F. Marques, "General region merging approaches based on information theory statistical measures," *ICIP*, pp. 3016–3019, 2008.

[2] F. Calderero and F. Marques, "General region merging based on first order Markov information theory statistical measures," *EUSIPCO, 2008*.

[3] D. Comaniciu, V. Ramesh, and P. Meer, "Kernel-based object tracking," *IEEE Trans. Patt. Anal. Mach. Intell.*, vol.25, no.5, pp.564-577, 2003.

[4] H. Xie, V. Andreu, and A. Ortega, "Quantization-based probabilistic feature modeling for kernel design in content-based image retrieval," *MIR'06*, pp. 23–32, 2006.

[5] P. Hall and E.J. Hannan, "On stochastic complexity and nonparametric density estimation," *Biometrika*, vol. 75, no. 4, pp. 705–714, 1988.

[6] Q. Wang, S.R. Kulkarni, and S. Verdu, "Divergence estimation of continuous distributions based on data-dependent partitions," *IEEE Trans. Inf. Theory*, vol. 51, no. 9, pp. 3064–3074, 2005.

[7] D.W. Scott and S.R. Sain, "Multi-dimensional density estimation," *Handbook of Statistics: Data Mining and Comput. Stats.*, vol. 23, 2004.

[8] M.B. Patel, J.J. Rodriguez, and A.F. Gmitro, "Effect of gray-level requantization on co-occurrence based texture analysis," *ICIP'08*, pp. 585–588, 2008.

[9] V. Vilaplana, F. Marques, and P. Salembier, "Binary partition trees for object detection," *IEEE Trans. Image Process.*, vol. 17, no. 11, pp. 2201–2216, 2008.

[10] J.S. Cardoso and L. Corte-Real, "Toward a generic evaluation of image segmentation," *IEEE Trans. Image Process.*, vol. 14, no. 11, pp. 1773-1782, 2005.