Robust image binarization with ensembles of thresholding algorithms

Farid Melgani
University of Trento
Department of Information and Communication Technologies
Via Sommarive, 14
I-38050 Trento, Italy
E-mail: melgani@dit.unitn.it

Abstract. *The effectiveness of a thresholding algorithm strongly depends on the image statistical characteristics. In a completely unsupervised context, this makes it difficult to choose the most appropriate algorithm to binarize a given image. This issue is considered through a novel thresholding strategy based on the fusion of an ensemble of different thresholding algorithms and formulated within a Markov random field (MRF) framework. The obtained experimental results suggest that in general the fusion of an ensemble of thresholding algorithms leads to a robust thresholding system, and in particular the proposed MRF strategy represents an effective solution to carry out the fusion process. © 2006 SPIE and IS&T.* [DOI: 10.1117/1.2194767]

1 Introduction

Image binarization is generally viewed as a low-level processing routine useful to extract objects from an image for their effective representation. The simplest approach to the binarization of gray-level images consists of selecting a global threshold in the image histogram to discriminate between object and background. In this context, numerous algorithms based on different mathematical approaches have been proposed to deal with the problem of the automatic selection of the best threshold value. They can be classified into two categories, namely, parametric and nonparametric algorithms. In the former category, the gray-level distribution of the "object" and "background" classes is assumed to follow a predefined statistical distribution that is typically of the Gaussian type.[1–3] Nonparametric techniques are based on the idea of finding the optimal global threshold through the optimization of a given criterion, which can be of a statistical type,[4,5] based on an entropy function,[6,7] or formulated under the fuzzy set theory.[8,9] A comprehensive overview of other image thresholding algorithms can be found in Refs. 10 and 11. Depending on the statistical characteristics of an image (e.g., the statistical distribution of the "object" and "background" classes, the degree of overlap between them, their prior probabilities), one thresholding algorithm may be more suitable than another.[10–12] However, in general, it is not trivial to choose the best algorithm for a given image, since the binarization process is unsupervised (i.e., no ground truth is available to guide the process).

A possible approach to overcome this problem is to fuse the results provided by an ensemble of different thresholding algorithms. In this way, it will be possible to exploit the peculiarities of the different thresholding algorithms synergetically, thus resulting in more robust final decisions than with a single thresholding algorithm. Note that the goal of the fusion is not to outperform the single-thresholding algorithm but to obtain accuracies comparable to that of the best single-thresholding algorithm independently of the image statistical characteristics. Such an approach has been studied extensively in the pattern recognition literature for the solution of challenging classification problems.[13–18] It has represented an important research topic and has been referred to in many ways, including classifier fusion, multiple classifier systems, and mixture of experts. From these works, it emerges that the fusion of an ensemble of classifiers can improve not only the robustness of the classification process but also the final classification accuracy. In practice, to increase the likelihood of obtaining better robustness and possibly better accuracy, two conditions are required: (1) diversity in the classifier ensemble and (2) an appropriate choice of fusion strategy.

The former is typically achieved by constructing ensembles of classifiers (1) based on different classification methodologies, (2) based on the same classification methodology but trained with different parameter values, (3) trained on different input feature spaces, or (4) manipulating different training sets by boosting[19,20] or bagging.[21] The latter will depend on the outputs provided by the classifiers, on their compatibility, and on whether training (labeled) samples are available for the fusion task. Among the most common fusion techniques, one can find linear and nonlinear fusion rules, such as the majority vote, the weighted majority vote, the product, the sum, the min, the max, and the median rules.[13,14,17,18,22] Other sophisticated fusion techniques worth mentioning are based on artificial neural networks,[15] on the dynamic selection of the classifier by the concept of local accuracy,[16] and on the fuzzy set theory.[22]

By contrast compared to the literature related to the classifier fusion, this attractive approach did not receive the attention it deserves in the image thresholding literature. Although a thresholding problem can be viewed as a clas-

Paper 05111RR received Jun. 19, 2005; revised manuscript received Nov. 15, 2005; accepted for publication Nov. 30, 2005; published online May 2, 2006.
1017-9909/2006/15(2)/023010/11/$22.00 © 2006 SPIE and IS&T.

sical binary classification problem, the implementation of a fusion strategy for thresholding algorithms is made difficult for two main reasons: (1) it should be carried out in a completely unsupervised way (i.e., with no labeled samples to represent the prior knowledge of the scene) and (2) the conceptual heterogeneity of thresholding algorithms leaves room only for a decision-level-based fusion because of the difficulty extracting compatible partial decision information that could be exploited in the fusion process. In this context, the aim of this paper is twofold. First, we propose to investigate the effectiveness of the fusion approach for the robust image thresholding problem through two classical fusion strategies based on the majority vote rule (MVR) and the weighted majority vote rule (WMVR), respectively. Second, we propose a novel fusion strategy formulated within a Markov random field (MRF) framework.

The rest of the paper is organized as follows. Section 2 presents the two investigated majority-based fusion strategies. Section 3 develops the proposed MRF fusion strategy. The three fusion strategies are experimentally compared in Sec. 4. Finally, Sec. 5 summarizes the main results of the paper.

2 Majority-Based Fusion Strategies

Let $X=\{x_{mn}: m=0,1,\ldots,M-1, n=0,1,\ldots,N-1\}$ be the original scalar $M\times N$ image with L possible gray levels ($x_{mn}=z, z\in\{0,1,\ldots,L-1\}$). Let us consider an ensemble of P different thresholding algorithms. Let T_i ($i=1,2,\ldots,P$) be the optimal threshold found by the i'th algorithm of the ensemble. Let $\boldsymbol{\omega}^i(x_{mn})=[\omega_1^i(x_{mn}),\omega_2^i(x_{mn})]$ be the binary decision vector of the i'th thresholding algorithm associated with pixel x_{mn} of the image and referring either to the "object" class or to the "background" class, such that

$$\boldsymbol{\omega}^i(x_{mn})=\begin{cases}[1,0] & \text{if } x_{mn}\leqslant T_i\\ [0,1] & \text{if } x_{mn}>T_i.\end{cases} \tag{1}$$

The goal of a fusion strategy based on the concept of majority is to construct the final global decision by consensus, i.e., by selecting the winning class that receives the largest number of favorable decisions. This can be implemented in two ways. One is based on the simple MVR. The other, called the WMVR, is an extension of the former and introduces the idea of weights to control the influence of each thresholding algorithm in the final global decision.

2.1 *MVR*

The MVR represents a fusion strategy, which is particularly attractive for its very simple implementation. In the context of the combination of classifiers, this rule has sometimes been found as successful as other more sophisticated combination strategies.[22,23] The MVR assigns the pixel x_{mn} to the label that obtains the highest number of votes among the ensemble of thresholding algorithms. In other words, a global decision vector $\boldsymbol{\Omega}(x_{mn})=[\Omega_1(x_{mn}),\Omega_2(x_{mn})]$ is derived by counting the votes from the members of the ensemble:

$$\boldsymbol{\Omega}(x_{mn})=\sum_{i=1}^{P}\boldsymbol{\omega}^i(x_{mn}). \tag{2}$$

The resulting thresholded image $Y=\{y_{mn}: m=0,1,\ldots,M-1, n=0,1,\ldots,N-1\}$ provided by the MVR will be a binary image in which each pixel y_{mn} takes on one of the following two labels:

$$y_{mn}=\begin{cases}\varphi_1 & \text{if } \Omega_1(x_{mn})\geqslant\Omega_2(x_{mn})\\ \varphi_2 & \text{if } \Omega_1(x_{mn})<\Omega_2(x_{mn}).\end{cases} \tag{3}$$

Typically, the labels φ_1 and φ_2, which refer to the "background" and "object" classes, are represented by the black and white colors (i.e., they take on digital values from the set {0, 255}).

2.2 *WMVR*

The underlying idea of the WMVR is to better control the influence of the members (experts) of the ensemble in the fusion process, since they can exhibit different accuracies. A simple way to carry out such a task is to assign a weight (degree of confidence) to the decision of each expert in the team.[14,15,22] According to the WMVR, the global decision vector $\boldsymbol{\Omega}(x_{mn})$ is obtained by summing the weighted decisions from the different experts:

$$\boldsymbol{\Omega}(x_{mn})=\sum_{i=1}^{P}\alpha^i(x_{mn})\cdot\boldsymbol{\omega}^i(x_{mn}), \tag{4}$$

where $\alpha^i(x_{mn})$ stands for the weight value associated with the i'th thresholding algorithm ($i=1,2,\ldots,P$) when applied to pixel x_{mn}. To generate the thresholded image Y, the decision rule expressed in Eq. (3) is adopted.

At this point, the main problem is to define the weight function $\alpha^i(\cdot)$. Several ways of determining the weights, such as those based on supervised linear or nonlinear optimization methods[14,15] and on the estimate of the individual classifier accuracies[22] are reported in the classifier fusion literature. In our case, the absence of *a priori* knowledge about the scene (i.e., unavailability of training samples) renders this task difficult. To this end, in this paper, we propose to exploit the threshold value found by each single-thresholding algorithm of the ensemble to derive a confidence measure based on the idea that the larger the difference (distance) between the pixel and threshold values, the higher the degree of confidence in the decision of the single-thresholding algorithm. A simple weight function $\alpha^i(\cdot)$ that satisfies such requirement is given by

$$\alpha^i(x_{mn})=1-\exp(-\gamma|x_{mn}-T_i|), \tag{5}$$

where γ is a real positive constant controlling the steepness of the weight function (Fig. 1). A general block diagram of the fusion strategy based on the WMVR is depicted in Fig. 2.

3 MRF Fusion Strategy

Because of the flexible and powerful stochastic framework they provide, MRFs have proved to be promising in many image processing areas, such as image restoration,[24] image and texture synthesis,[25] segmentation,[26] and classification.[27]

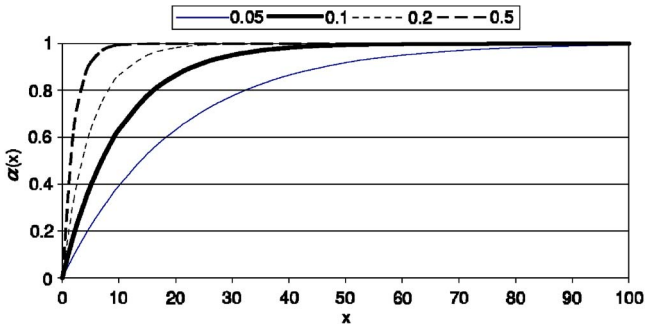


Fig. 1 Plot of the weight function $\alpha(\cdot)$ for different values of the steepness parameter γ.

In particular, in image classification, MRFs have been found to be an effective way of fusing sources of information of a different nature in the classification scheme.[28,29] The Markovian approach makes it possible to reduce the fusion task into a problem of minimizing a total energy function, which, in addition to an energy function describing the local spatial properties of the image, aggregates the energy functions associated with the considered information sources.

In this paper, we propose a novel strategy that exploits the attractive properties of MRFs to combine the results provided by an ensemble of thresholding algorithms.

3.1 *MRF Fusion Model Formulation*

Let $A_i(i=1,2,\ldots,P)$ be the thresholded image generated by the i'th thresholding algorithm of the ensemble. Since a thresholding problem can be viewed as a binary classification problem where each pixel (m,n) is assigned to a label $y_{mn} \in \{\varphi_1, \varphi_2\}$, the optimal classification Y^* of all the pixels of the original image X, given the thresholded images $A_i(i=1,2,\ldots,P)$, can be performed by applying the maximum *a posteriori* probability (MAP) decision criterion:

$$P(Y^*|A_1,A_2,\ldots,A_P) = \max_{Y}\{P|A_1,A_2,\ldots,A_P\}. \quad (6)$$

By adopting the MRF approach, one can greatly simplify the complexity of this maximization problem by passing from a global model to a model of the local image properties. The latter is defined both in terms of the potential function of individual pixels and of the interactions among pixels in appropriate neighborhoods. The combination of the MAP method with the MRF modeling makes our binary classification task equivalent to the minimization of a total energy function U_T expressed in the following relationship:

$$P(Y|A_1,A_2,\ldots,A_P) = \frac{1}{Z}\exp[-U_T(Y|A_1,A_2,\ldots,A_P)], \quad (7)$$

where Z is a normalizing constant.

Under the Markovian approach, the total energy function $U_T(\cdot)$ can be rewritten in terms of local energy functions U_{mn} using the concept of neighborhood:[30]

$$U_T(Y|A_1,A_2,\ldots,A_P) = \sum_{m=0}^{M-1}\sum_{n=0}^{N-1} U_{mn}, \quad (8)$$

with

$$U_{mn} = U[y_{mn}, Y^S(m,n), A_1^S(m,n), A_2^S(m,n), \ldots, A_P^S(m,n)], \quad (9)$$

where $Y^S(m,n)$ and $A_i^S(m,n)$ stand for the set of labels of the pixels of the image Y and the images A_i $(i=1,2,\ldots,P)$, respectively, in a predefined neighborhood system S associated with pixel (m,n).

The minimization of Eq. (8) can be carried out by means of different optimal or suboptimal methods; the most popular are simulated annealing (SA), the maximizer of posterior marginals (MPM), the iterated conditional modes (ICM) algorithms,[30] and the graph cuts methods.[31] In this paper, the ICM algorithm is adopted since it represents a simple and computationally moderate solution to optimize the MRF-MAP estimates, for it converges to a local, but usually good, minimum of the energy function. The ICM consists of minimizing iteratively the total energy function $U_T(\cdot)$ through a pixel-based scheme until convergence is

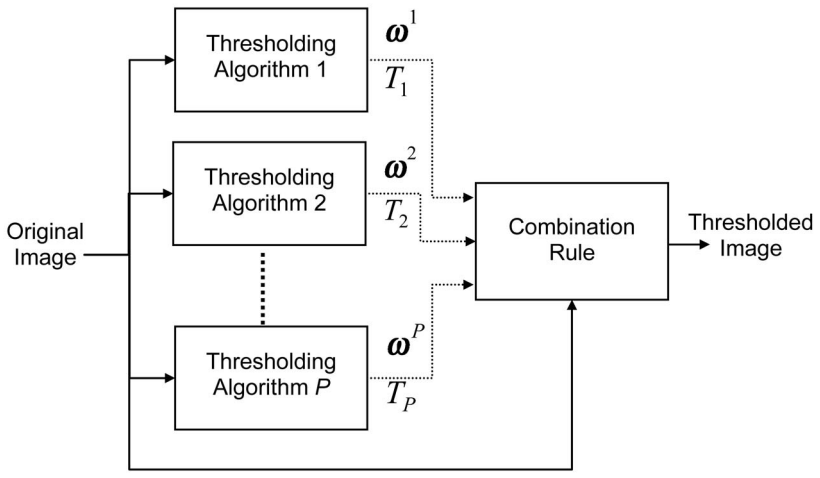


Fig. 2 General block diagram of the fusion strategy based on the WMVR.

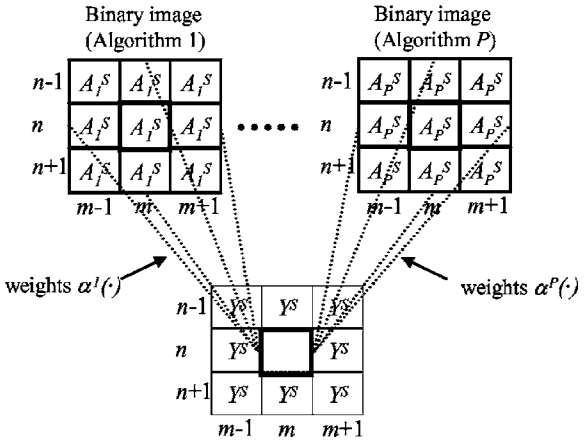


Fig. 3 Neighborhood system associated to pixel (m,n) of the image; Y^S and A_i^S stand for the spatial and interimage neighborhoods, respectively.

reached (i.e., where the pixel labels do not change much). In other words, the optimization process is reduced to the iterative minimization of the local energy function U_{mn} associated with each pixel (m,n). As the true set of labels $Y^S(m,n)$ in Eq. (9) is unknown, at each iteration the estimate of $Y^S(m,n)$ obtained at the previous iteration is used to generate a new estimate of the label set Y.

At this point, the first problem to deal with is the decomposition of the local energy function U_{mn}. This depends on two kinds of sources of contextual information, which contribute to the optimization process. They are (1) the spatial contextual information source, which defines the spatial correlation in image Y between the label of pixel (m,n) and the labels of its neighbors, and (2) the interimage information sources, which express the relationship between the image Y and each of the thresholded images A_i $(i=1,2,\ldots,P)$. Similarly to what is done in Refs. 28 and 29 in the context of multisource classification, for the sake of simplicity, it is assumed that the contributions from these sources of information are separable and additive. Accordingly, the local energy function U_{mn} to be minimized for the pixel (m,n) can be written as follows:

$$U_{mn} = \beta_{\mathrm{SP}} \cdot U_{\mathrm{SP}}[y_{mn}, Y^S(m,n)] + \sum_{i=1}^{P} \beta_i \cdot U_{\mathrm{II}}[y_{mn}, A_i^S(m,n)], \tag{10}$$

where $U_{\mathrm{SP}}(\cdot)$ and $U_{\mathrm{II}}(\cdot)$ refer to the spatial and interimage energy functions, respectively, while β_{SP} and β_i $(i=1,2,\ldots,P)$ represent the spatial and interimage parameters, respectively.

3.2 *Energy Functions*

The neighborhood system $S=Y^S \cup A_1^S \cup \cdots \cup A_P^S$ adopted to define the two kinds of energy functions required to compute the local energy function in Eq. (10) is based on a second-order neighborhood (Fig. 3). On the basis of this neighborhood system, the spatial energy function can be expressed as:[30]

$$U_{\mathrm{SP}}[y_{mn}, Y^S(m,n)] = -\sum_{y_{pq} \in Y^S(m,n)} I(y_{mn}, y_{pq}), \tag{11}$$

where $I(.,.)$ is the indicator function, which enables us to count the number of occurrences of y_{mn} in Y^S (the spatial part of S) and is defined as

$$I(y_{mn}, y_{pq}) = \begin{cases} 1, & \text{if } y_{mn} = y_{pq} \\ 0, & \text{otherwise.} \end{cases} \tag{12}$$

In a similar way as in the spatial correlation, we define the correlation between the image Y and the images A_i $(i=1,2,\ldots,P)$ and, accordingly, the interimage energy function as follows:

$$U_{\mathrm{II}}[y_{mn}, A_i^S(m,n)] = -\sum_{A_i(p,q) \in A_i^S(m,n)} \alpha^i(x_{pq}) \cdot I[y_{mn}, A_i(p,q)]. \tag{13}$$

The use of the weight function $\alpha^i(\cdot)$, as defined in Eq. (5), aims at controlling, during the fusion process, the effect of unreliable decisions at the pixel level that can be incurred by the thresholding algorithms. The possible misleading effects of the latter are further controlled at a global (image) level through the interimage parameters β_i $(i=1,2,\ldots,P)$, which are computed as follows:

$$\beta_i = \exp(-\gamma|\bar{T} - T_i|), \tag{14}$$

where $\bar{T}$ is the average threshold value:

$$\bar{T} = \frac{1}{P}\sum_{i=1}^{P} T_i. \tag{15}$$

Accordingly, with this global weighting mechanism, a thresholding algorithm is penalized if it exhibits a threshold value that is statistically incompatible with those of the ensemble.

3.3 *Algorithm*

The algorithm of the proposed MRF fusion strategy can be summarized as follows:

3.3.1 *Initialization step*

1. Apply each thresholding algorithm of the ensemble on image X to generate the set of thresholded images A_i $(i=1,2,\ldots,P)$.
2. Initialize Y by minimizing for each pixel (m,n) the local energy function U_{mn} defined in Eq. (10) without the spatial energy term (i.e., by setting $\beta_{\mathrm{SP}}=0$).

3.3.2 *K'th iteration*

Update Y by minimizing for each pixel (m,n) the local energy function U_{mn} defined in Eq. (10) including the spatial energy term (i.e., by setting $\beta_{\mathrm{SP}} \neq 0$).

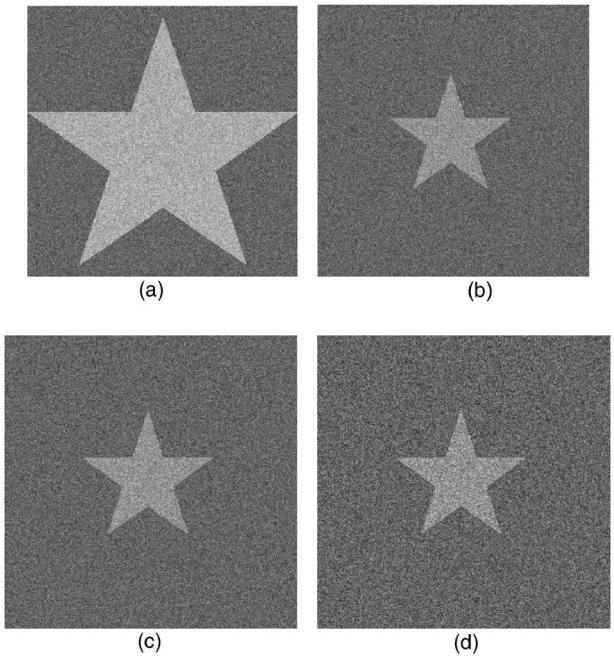


Fig. 4 Simulated test images used to assess the accuracy of the different thresholding algorithms and fusion strategies: (a) image 1 (μ_1=100, σ_1=20, P_1=69.3%; μ_2=175, σ_2=20, P_2=30.7%); (b) image 2 (μ_1=100, σ_1=20, P_1=93.7%; μ_2=150, σ_2=20, P_2=6.3%); (c) image 3 obtained from image 2 after addition of uniform noise (PSNR=31.2 dB and RMSE=7); image 4 obtained from image 2 after addition of uniform noise (PSNR=22.9 dB and RMSE=18.2).

3.3.3 *Stop criterion*

Repeat step 2 $K_{\max}$ times or until the number of different labels in Y computed over the last two iterations becomes very small.

4 Experimental Results

4.1 *Data Set Description*

An assessment of the performance of the different thresholding strategies was carried out on the basis of eight 8-bit gray-level images. By simulating different statistical conditions, the first group of four images aims at testing the sensitivity of the strategies to the problems of (1) the balance between the "background" and "object" classes (i.e., the problem of prior probabilities), (2) overlap between these two classes, and (3) noise corruption. In these four simulated images, samples of both the "background" and "object" classes were drawn from two different normal distributions $N(\mu_i, \sigma_i)$ with prior probabilities P_i (i=1 for "background" and i=2 for "object"), where the two parameters μ_i and σ_i stand for the mean and the standard deviation of the distribution, respectively (see Figs. 4 and 5). In particular, the first simulated image (image 1) refers to a noiseless image where the two classes are well separated and not strongly disproportioned. In the second noiseless image (image 2), unlike image 1, the two classes were relatively strongly overlapped and unbalanced. The third and fourth images (images 3 and 4) were generated from image 2 after adding uniform noise with increasing power. The

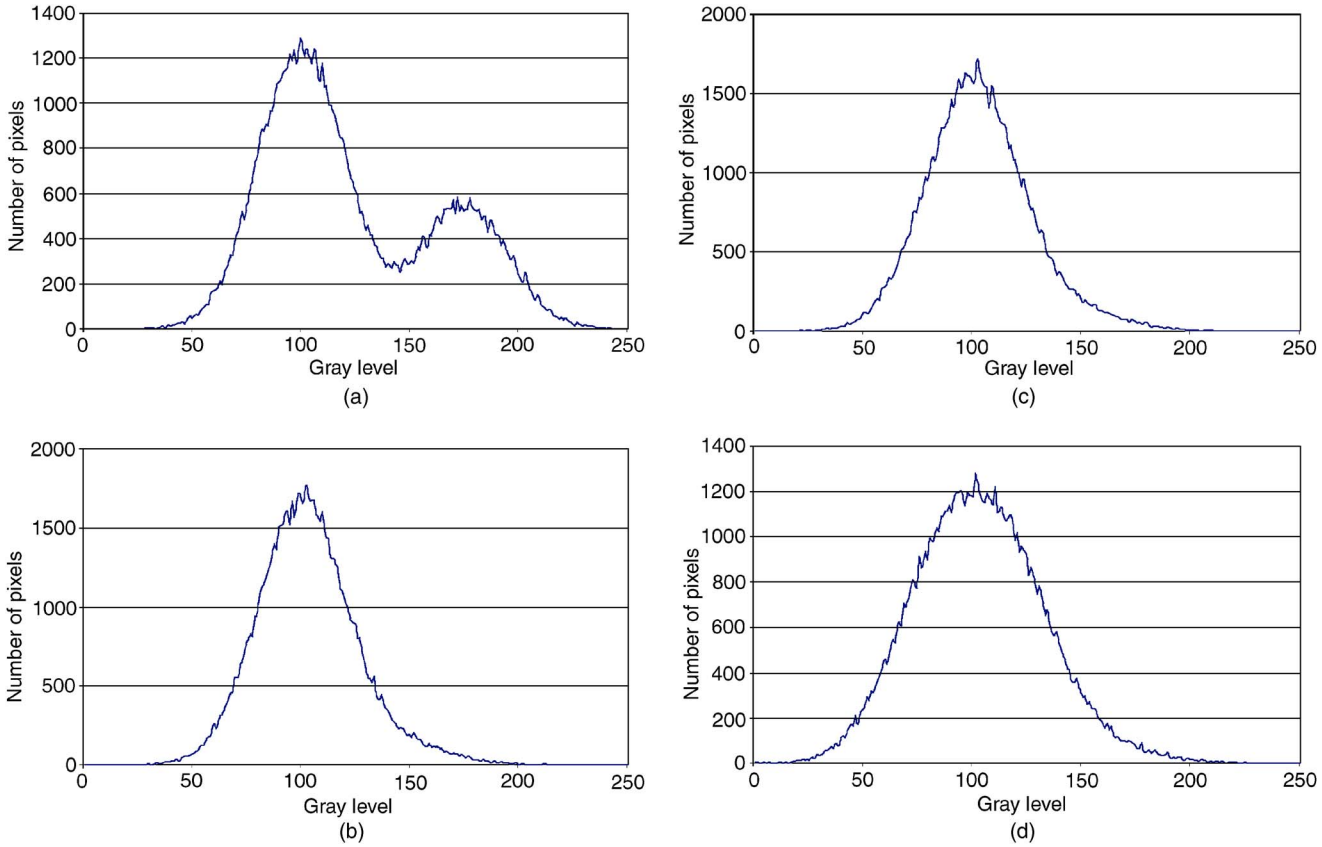


Fig. 5 Histograms of the four simulated test images: (a) image 1, (b) image 2, (c) image 3, and (d) image 4.

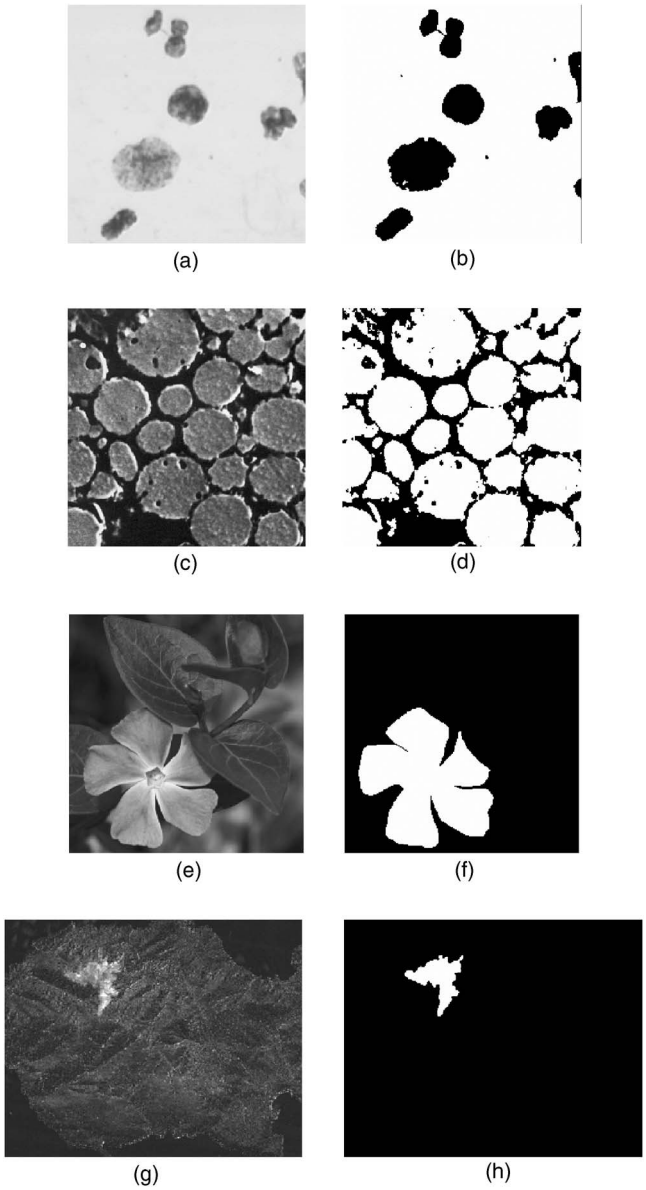


Fig. 6 Real test images used to assess the accuracy of the different thresholding algorithms and fusion strategies: (a) original cell image, (b) ground-truth of cell image, (c) original material structure image, (d) ground-truth of material structure image, (e) original flower image; (f) ground-truth of flower image, (g) original remote-sensing image, and (h) ground-truth of remote-sensing image.

effect of noise was quantified in terms of two standard image quality measures that are the peak signal-to-noise ratio (PSNR) and the root mean squared error (RMSE) measures.

The second group of four images represents real images related to different application fields (see Fig. 6). In greater detail, the first image of this group (image 5) refers to a cell image acquired by light microscopy. The second (image 6) represents a light microscope image of a material structure. The third (image 7) is a flower image while the fourth (image 8) is generated from two multitemporal multispectral remote-sensing images acquired over the Italian island of Elba to map a forest fire that occurred in 1994. The corresponding ground-truth images, which were obtained by expert's photointerpretation, are also shown in Fig. 6. Note how the histograms of the four real test images provided in Fig. 7 clearly illustrate their different statistical characteristics, which are due to their different origin.

The accuracies of the three fusion strategies and of each single-thresholding algorithm of the ensemble were estimated in terms of rates of errors and of false and missed alarms. The error rate, which represents a global error criterion, is defined as the ratio of the total number of misclassified pixels to the total number of pixels. The false and missed alarms rates are more specific error criteria since they refer to the ratios of the number of misclassified background and of misclassified object pixels to the total number of background and of object pixels, respectively.

4.2 *Results of the Thresholding Algorithms*

To construct the ensemble, four different popular thresholding algorithms were considered. These are the Kittler and Illingworth,[2] the Otsu,[4] the Kapur *et al.*,[6] and the Huang and Wang[9] algorithms. The four algorithms were run for each of the eight test images so as to provide the threshold values T_i, the single decision vectors $\boldsymbol{\omega}^i$ $(i=1,\ldots,4)$ for each image pixel and, accordingly, the thresholded images A_i $(i=1,\ldots,4)$. These outputs were exploited in the successive step by the described fusion strategies to produce a global thresholded image Y.

All the results obtained by the single thresholding algorithms in terms of threshold values, of error rate (ER) and of false and missed alarms rates (FA and MA, respectively) are reported in Tables 1 and 2 for the simulated and real images, respectively.

Concerning the simulated images, as expected, all the four thresholding algorithms could achieve a good accuracy on image 1 because of the high discriminability between "background" and "object" classes. Image 1 thus provides a fusion scenario of decisions of experts in almost complete agreement. On the contrary, image 2 has created strong difficulties to the Otsu and the Huang and Wang algorithms, which exhibited poor error rates of 37.71 and 45.21%, respectively, but also to the Kittler and Illingworth algorithm due to its high MA rate (equal to 72.93%). The Kapur *et al.* algorithm proved to be the best in such a situation of strong unbalancing and overlap between "background" and "object" classes. The two algorithms that showed a higher sensitivity to the addition of noise in image 2 are the Kittler and Illingworth and the Kapur *et al.* algorithms. Their threshold values moved from 162 (image 2) to 168 (image 3) and then to 202 (image 4), and from 140 (image 2) to 141 (image 3) and then to 156 (image 4), respectively. This had the consequence of increasing significantly their MA rate from 72.93 (image 2) to 97.37% (image 4) and from 31.14 (image 2) to 58.89% (image 4), respectively. They remain, however, the best algorithms of the ensemble in terms of error rate for both images 3 and 4. These last images, in addition to image 2, represent critical but interesting fusion scenarios of decisions of experts in strong disagreement.

Moving to the results obtained on the set of real images, for image 5, the error rate values exhibited by the thresholding algorithms are relatively close to each other. They range from 0.66% for the Kapur *et al.* algorithm to 3.02% for the Otsu algorithm. Such good agreement between the

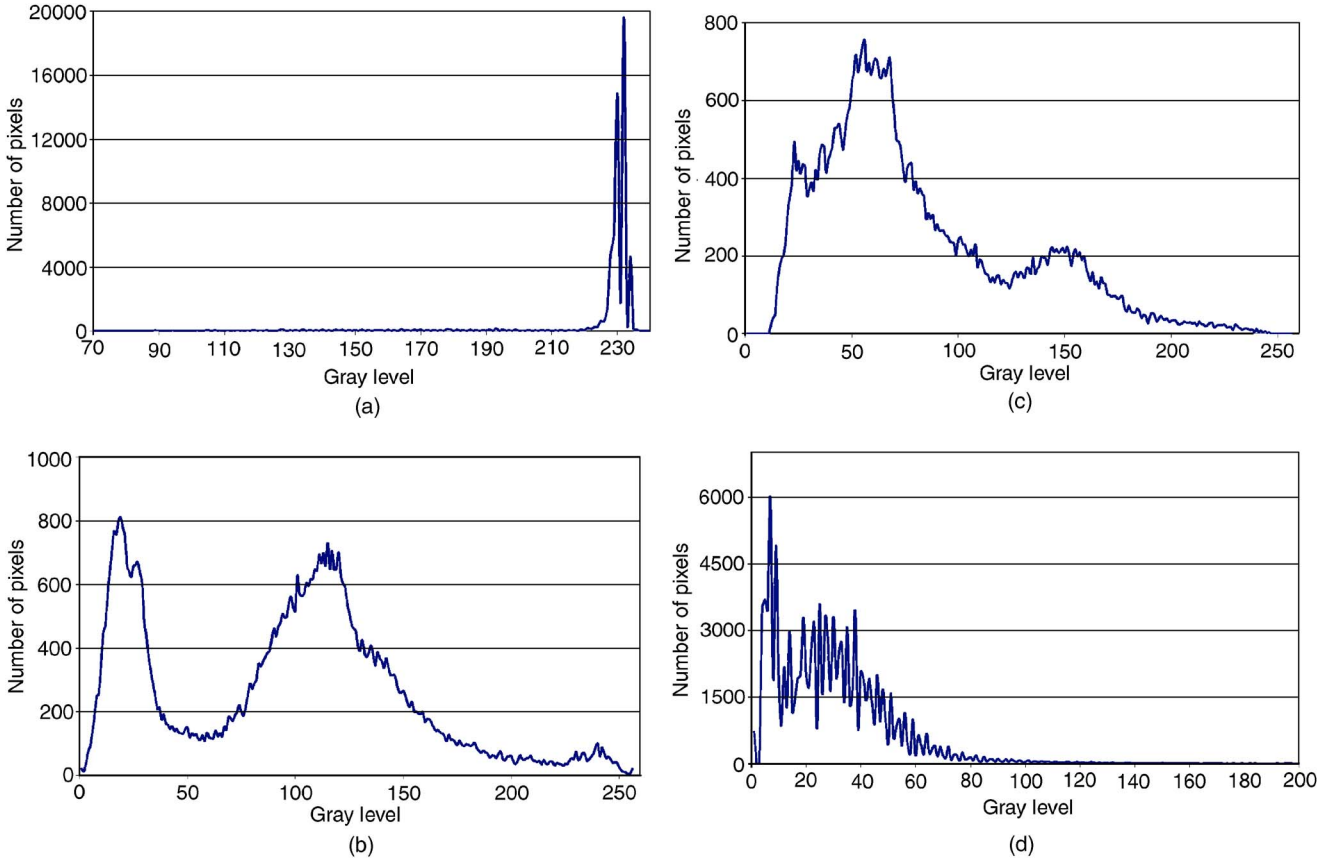


Fig. 7 Histograms of the four real test images: (a) image 5, (b) image 6, (c) image 7, and (d) image 8.

different algorithms is confirmed by a visual inspection of their corresponding thresholded images [see Figs. 8(a)–8(d)]. The same tendency can be observed for image 7, where the best algorithm was the Kittler and Illingworth algorithm with an error rate of 5.97%, while the worst was the Huang and Wang algorithm with an error rate of 9.22%. For the remaining two images, i.e., images 6 and 8, the situation is completely different since the ensemble includes at least one poor thresholding algorithm. This situation, which also characterizes images 2, 3, and 4, is particularly useful to test the robustness of the different fusion strategies. For image 6, the best thresholding algorithm was the Huang and Wang algorithm, which found the best possible threshold value corresponding to an error rate of 0%.

Table 1 Threshold values (T), error rate (ER), and false and missed alarms rates (FA and MA, respectively) in percent obtained on the four simulated test images by the different thresholding algorithms and fusion strategies.

	Image 1				Image 2				Image 3				Image 4			
Method	T	ER (%)	FA (%)	MA (%)	T	ER (%)	FA (%)	MA (%)	T	ER (%)	FA (%)	MA (%)	T	ER (%)	FA (%)	MA (%)
Kittler and Illingworth	142	**2.65**	1.76	4.66	162	**4.71**	0.09	72.93	168	**5.12**	0.05	80.22	202	**6.16**	0	97.37
Otsu	135	**3.41**	3.96	2.15	105	**37.71**	40.26	1.56	105	**38.11**	40.55	1.95	103	**42.86**	45.45	4.60
Kapur *et al.*	131	**4.61**	6.09	1.27	140	**4.04**	2.20	31.14	141	**4.60**	2.62	33.88	156	**5.28**	1.64	58.89
Huang and Wang	137	**3.02**	3.17	2.69	101	**45.21**	48.21	0.90	100	**46.96**	50.04	1.26	101	**45.58**	48.39	3.93
MVR	—	**3.04**	3.20	2.70	—	**19.83**	19.98	17.68	—	**20.15**	20.27	18.38	—	**23.16**	22.47	33.44
WMVR	—	**3.02**	3.17	2.69	—	**10.58**	10.56	10.92	—	**11.16**	10.99	13.60	—	**12.18**	11.16	27.26
MRF	—	**0.08**	0.08	0.10	—	**0.14**	0.04	1.65	—	**0.16**	0.03	1.97	—	**0.46**	0	7.18

Table 2 Threshold values (T), error rate (ER), and false and missed alarms rates (FA and MA, respectively) in percent obtained on the four real test images by the different thresholding algorithms and fusion strategies.

	Image 5				Image 6				Image 7				Image 8			
Method	T	ER (%)	FA (%)	MA (%)	T	ER (%)	FA (%)	MA (%)	T	ER (%)	FA (%)	MA (%)	T	ER (%)	FA (%)	MA (%)
Kittler and Illingworth	222	**1.53**	1.75	0	38	**3.13**	0	10.88	110	**5.97**	5.59	7.40	77	**1.58**	1.29	15.45
Otsu	181	**3.02**	0	24	79	**6.57**	9.23	0	100	**8.55**	9.73	4.07	35	**30.29**	30.91	0.21
Kapur *et al.*	217	**0.66**	0.75	0	162	**61.80**	86.78	0	107	**6.52**	6.63	6.07	94	**1.06**	0.35	35.67
Huang and Wang	196	**1.32**	0	10.51	52	**0**	0	0	98	**9.22**	10.66	3.69	22	**56.72**	57.89	0
MVR	—	**0.99**	0.41	5.08	—	**3.61**	5.07	0	—	**7.25**	7.74	5.39	—	**15.15**	15.30	7.79
WMVR	—	**0.36**	0	2.86	—	**2.86**	4.02	0	—	**7.29**	7.85	5.15	—	**6.39**	6.47	2.49
MRF	—	**0.35**	0.03	2.58	—	**2.89**	3.85	0.52	—	**6.50**	6.94	4.81	—	**2.91**	2.24	0

The Kapur *et al.* algorithm, which proved very accurate for images 5 and 7, failed completely this time (error rate equal to 61.8%). Image 8 illustrates another critical fusion scenario where the ensemble includes two poor algorithms (the Huang and Wang and the Otsu algorithms with error rates of 56.72 and 30.29%, respectively) and two accurate algorithms (the Kapur *et al.* and the Kittler and Illingworth algorithms with error rates of 1.06 and 1.58%, respectively).

In general, these experimental results confirm that the choice of a thresholding algorithm strongly depends on the image to be processed. For one image, an algorithm may appear the best, while it may fail completely for another. However, one would expect such a problem to be overcome by exploiting the synergies between the different algorithms through a proper fusion strategy.

4.3 *Results of the Fusion Strategies*

For the two WMVR and MRF fusion strategies, the steepness constant γ of the weight function defined in Eq. (5) was set to a value of 0.1, which generates a confidence degree of 90% for a difference value of around 25 between the threshold value and the pixel gray level. In addition, for the MRF strategy, it was needed to set the spatial parameter β_{SP}. Experiments were carried out by varying the value of such parameter from 0.5 to 2. The obtained results did not change significantly for the eight considered images, suggesting that the setting of such parameter is not critical. The detailed results reported in the following refer all to $\beta_{SP}=1$.

One of the problems of the MVR is that it requires the number of thresholding algorithms to be odd to avoid ties. In these experiments, since P is even and no rejection decision is allowed, the assignment of the pixel was done randomly to one of the two labels in the event of ties. One of the methodological advantages of the WMVR compared to the MVR is that it is hardly affected by such a problem since the global decision vector $\mathbf{\Omega}(x_{mn})$ associated with each pixel of coordinates (m,n) is made up of real-valued components.

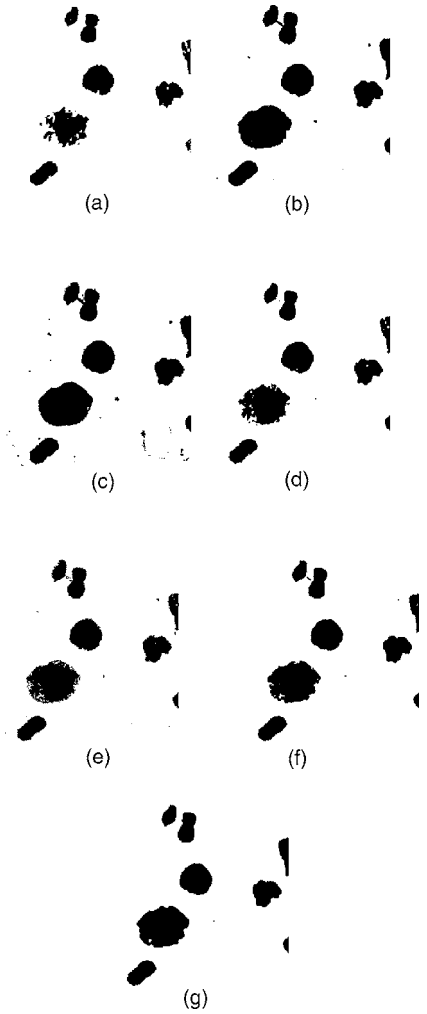


Fig. 8 Binary maps obtained by the different thresholding algorithms and fusion strategies for the first light microscope image (image 5): (a) Otsu algorithm, (b) Kapur *et al.* algorithm, (c) Kittler and Illingworth algorithm, (d) Huang and Wang algorithm, (e) MVR fusion strategy, (f) WMVR fusion strategy, and (g) MRF fusion strategy.

In general, the results achieved by the three fusion strategies for images 1, 5, 6, and 7 were very satisfactory. For images 2, 3, 4 and 8, which represent the most critical fusion scenarios, significant differences between the different strategies can be observed.

In greater detail, for images 1 and 7, which translate two fusion scenarios of decisions of experts in strong agreement, we can observe that a better result was achieved by the MRF strategy compared to the two other fusion strategies. Its ER is very close to that exhibited by the Kittler and Illingworth algorithm, which represents the best thresholding algorithm of the two ensembles associated with these two images (ERs equal to 0.08 against 2.65% and 6.50 against 5.97% for images 1 and 7, respectively). The two other fusion strategies, i.e., the MVR and the WMVR strategies could improve slightly the ER with respect to the average ER of the ensemble. In these two images, it appears clearly that, when experts are in strong agreement and accordingly correlated, it is very likely that the fusion of their decisions will not help to improve significantly the results.

Image 5 can be seen as an example of fusion of thresholding algorithms relatively accurate and weakly correlated since their threshold values vary in a wide interval that goes from 181 to 222. For this case of fusion, the WMVR and MRF strategies led to almost half the ER compared to the result achieved by the best thresholding algorithm of the ensemble (ERs equal to 0.36 and 0.35% for the WMVR and MRF, respectively, against 0.66 for the Kapur *et al.* algorithm). The simplest fusion strategy, i.e., the MVR, incurred an ER that was close to that of the best thresholding algorithm of the ensemble (0.99 against 0.66%). The thresholded maps obtained by the four thresholding algorithms and the three fusion strategies are depicted in Fig. 8. A visual inspection of the maps and their comparison with the related ground-truth shown in Fig. 6(b) enables us to confirm (1) the promising results of the WMVR and MRF strategies and (2) that, as expected theoretically, the MRF strategy leads to a map that has a higher likelihood to be cleared of isolated labeling.

Image 6 represents a good robustness test for the three investigated fusion strategies, since as mentioned previously the ensemble includes a poor thresholding algorithm—that is, the Kapur *et al.* algorithm. Despite such a misleading source of information (characterized by an ER of 61.8%), the three strategies proved robust. In particular, thanks to the WMVR and MRF strategies, a result relatively close to that of the best thresholding algorithm and better than that of the second-best algorithm of the ensemble was achieved (ERs of 2.86 and 2.89% against 0 and 3.13%, respectively).

Another interesting robustness test is provided by images 2, 3, 4, and 8. The corresponding ensembles of thresholding algorithms seem difficult to handle because of the strong disagreement between their members (at most two accurate against at least two poor algorithms). Despite this constraint, the MRF strategy proved particularly effective in all four images. For instance, for the fourth simulated and real images (i.e., images 4 and 8), it enabled us to reach an ER very close to the ER incurred by the best single thresholding algorithm, i.e., the Kapur *et al.* algorithm (0.46 against 5.28% and 2.19 against 1.06% for images 4 and 8, respectively). The two other fusion strategies proved less robust than the MRF strategy since they could not handle well such a critical situation. Indeed, the WMVR incurred ERs of 12.18 and 6.39%, while the MVR completely failed with ERs of 23.16 and 15.15% for images 4 and 8, respectively. Figures 9(a)–9(d) and 10(a)–10(d) visually confirm the difficult task assigned to the three fusion

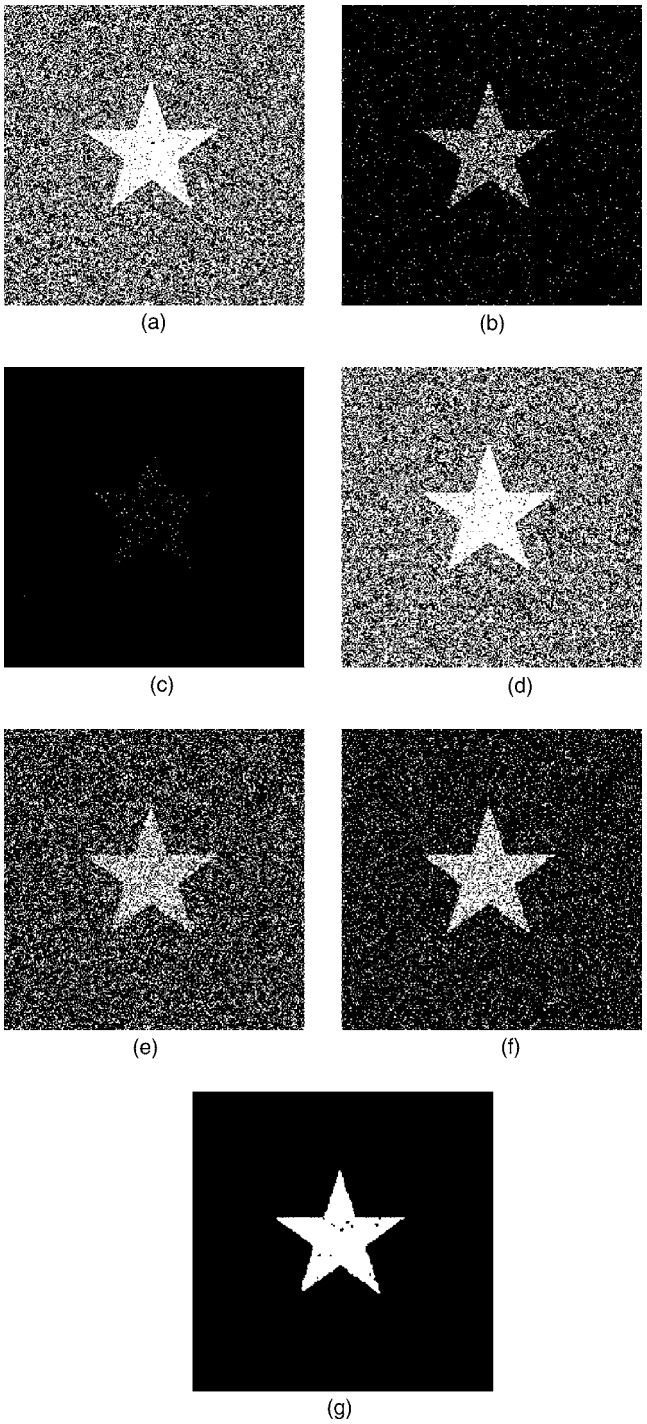


Fig. 9 Binary maps obtained by the different thresholding algorithms and fusion strategies for the fourth simulated image (image 4): (a) Otsu algorithm, (b) Kapur *et al.* algorithm, (c) Kittler and Illingworth algorithm, (d) Huang and Wang algorithm,(e) MVR fusion strategy, (f) WMVR fusion strategy, and (g) MRF fusion strategy.

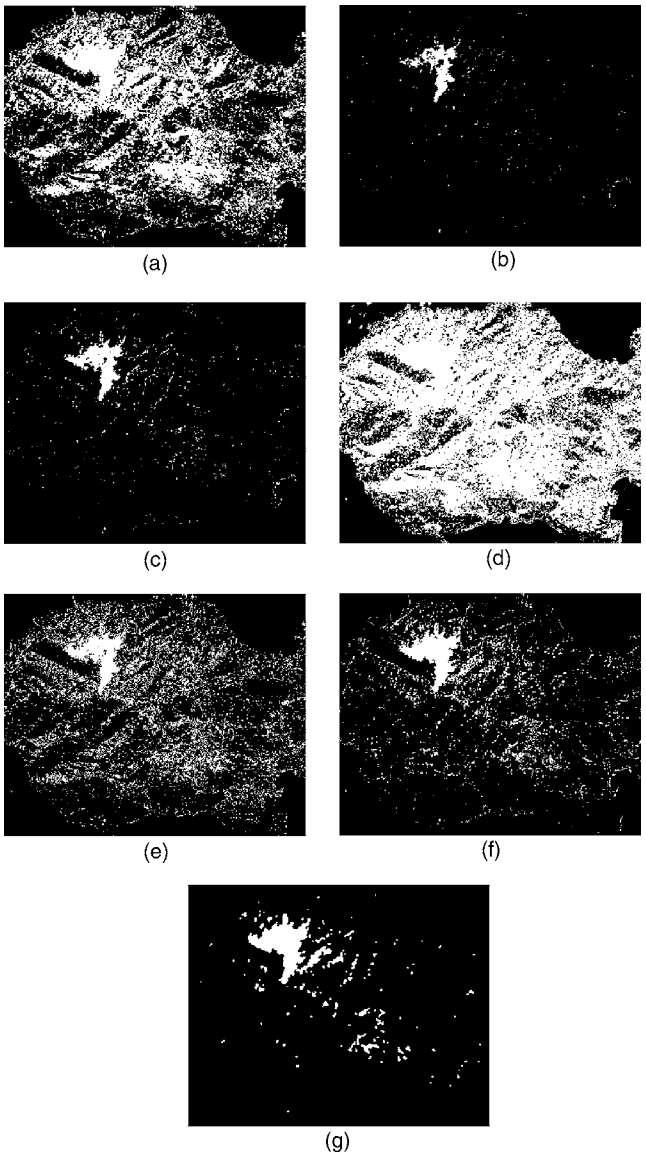


Fig. 10 Binary maps obtained by the different thresholding algorithms and fusion strategies for the remote-sensing image (image 8): (a) Otsu algorithm,(b) Kapur *et al.* algorithm, (c) Kittler and Illingworth algorithm, (d) Huang and Wang algorithm, (e) MVR fusion strategy, (f) WMVR fusion strategy, and (g) MRF fusion strategy.

strategies. Indeed, for both images 4 and 8, the thresholded maps associated with the Otsu and the Huang and Wang algorithms contain an excessive number of false alarms. Moreover, from image 4, the Kittler and Illingworth algorithm generated a binary map that suffers from a substantial number of missed alarms. Despite this difficult situation raised by both images 4 and 8, the MRF fusion strategy provided binary maps very close to the ground-truth maps with a number of FAs and MAs, which is much less significant than what was obtained by the MVR and WMVR strategies [see Figs. 9(e)–9(g) and 10(e)–10(g)].

Figure 11 plots two typical behaviors of the ER of the MRF fusion strategy after each processing iteration. For image 5, one can observe that the ER is stable. This suggests that the spatial information source exploited by the MRF fusion strategy appeared useless for this image. This

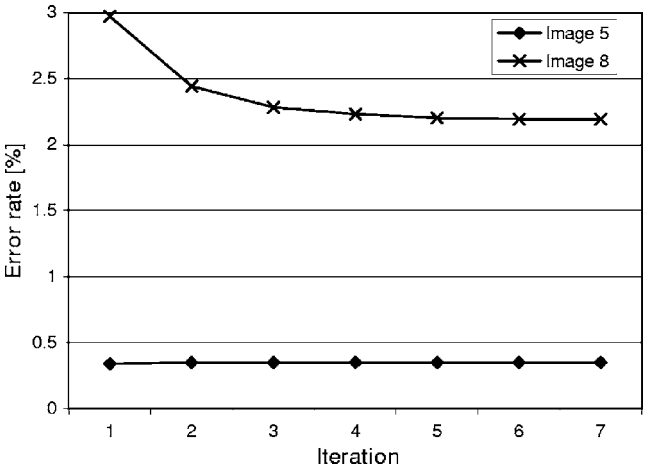


Fig. 11 Behavior of the ER incurred by the MRF fusion strategy versus the number of iterations for images 5 and 8.

can be explained by two facts: (1) the maps generated by the individual thresholding algorithms have already reached a significant degree of spatial homogeneity [see Fig. 8(a)–8(d)], and (2) some spatial information is indirectly conveyed by the interimage energy function (used to initialize the MRF strategy) on account of the adopted neighborhood system (see Fig. 3). The combination of these two facts left no room for the spatial energy function to reduce the ER further after the initialization iteration. For image 8, which is characterized by the presence of a more significant amount of noise due to its acquisition mode, the benefit of the spatial information source is clearly illustrated by Fig. 11. Indeed, a gain of 0.79% in ER is obtained on passing from the initialization iteration to the seventh iteration. This enabled us to reduce significantly the errors in the fusion process involved by the misleading decisions of the Otsu and the Huang and Wang thresholding algorithms.

5 Conclusions

The experimental results confirm that the choice of a thresholding algorithm strongly depends on the image to be processed. An algorithm may appear the best one for one image, while it may fail completely for another. A possible solution to this issue is represented by a thresholding approach based on the fusion of an ensemble of different thresholding algorithms. In this paper, to carry out the combination task, a novel fusion strategy based on MRFs was proposed and compared with other two strategies inspired by techniques used in multiple classifier systems.

In general, the obtained results suggest that with the combination of an ensemble of different thresholding algorithms, it is possible to capture the best peculiarities to achieve robust thresholding. This is shown by the fact that it leads to results often close to those of the best single-thresholding algorithm of the ensemble independently of image typology.

In greater detail, among the three investigated fusion strategies, the proposed MRF strategy proved to be the most effective because of (1) its natural capability to integrate the spatial contextual information in the fusion model and (2) a weighting mechanism implemented at both the pixel and image level to handle the reliability of the results

provided by each thresholding algorithm making up the considered ensemble. These two properties render it more robust especially to critical situations where no clear consensus can be reached between the thresholding algorithms because of the complexity of the original image. Such situations are typical when the "background" and "object" classes in an image are strongly unbalanced, overlapped, or affected by the presence of noise.

Note that the estimation of the pixel and image-level weights used in the proposed MRF fusion model is constrained by the unsupervised nature of the thresholding problem. Although more complex weight functions could be derived for this purpose, in this paper, we proposed a solution that leads to an MRF fusion strategy that is conceptually simple, easy to implement, and fast to run and has a high performance. A comparison between the two other MVR and WMVR strategies confirms the valuable role of the proposed simple weighting mechanism in improving the quality of the final global decision.

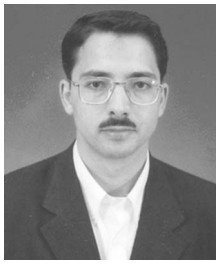
Finally, note that the fusion concept conveyed by the explored strategies is not limited to merge binarization algorithms based on global thresholding, as done in this work, but can be extended with opportune adaptations to the combination of other kinds of binarization algorithms such as adaptive thresholding and snake-based algorithms.

Acknowledgment

The author wishes to thank Dr. M. Sezgin (Tübitak Marmara Research Center, Turkey) for providing the two microscope images and their corresponding ground-truths used in the experiments.

References

1. J. S. Weszka and A. Rosenfeld, "Histogram modification for threshold selection," *IEEE Trans. Syst. Man Cybern.* **9**, 38–52 (1979).
2. J. Kittler and J. Illingworth, "Minimum error thresholding," *Pattern Recogn.* **19**, 41–47 (1986).
3. W. Snyder, G. Bilbro, A. Logenthiran, and S. Rajala, "Optimal thresholding—a new approach," *Pattern Recogn. Lett.* **11**, 803–809 (1990).
4. N. Otsu, "A threshold selection method from gray-level histogram," *IEEE Trans. Syst. Man Cybern.* **9**, 62–66 (1979).
5. W. Tsai, "Moment-preserving thresholding: a new approach," *Comput. Vis. Graph. Image Process.* **29**, 377–393 (1985).
6. J. N. Kapur, P. K. Sahoo, and A. K. C. Wong, "A new method for gray-level picture thresholding using the entropy of the histogram," *Comput. Vis. Graph. Image Process.* **29**, 273–285 (1985).
7. A. Brink, "Maximum entropy segmentation based on the autocorrelation function of the image histogram," *J. Comput. Inf. Technol.* **2**, 77–85 (1994).
8. S. K. Pal, R. A. King, and A. A. Hashim, "Automatic gray level thresholding through index of fuzziness and entropy," *Pattern Recogn. Lett.* **1**, 141–146 (1983).
9. L. K. Huang and M. J. Wang, "Image thresholding by minimizing the measures of fuzziness," *Pattern Recogn.* **28**, 41–51 (1995).
10. N. R. Pal and S. K. Pal, "A review on image segmentation techniques," *Pattern Recogn.* **26**, 1277–1294 (1993).
11. M. Sezgin and B. Sankur, "Survey over image thresholding techniques and quantitative performance evaluation," *J. Electron. Imaging* **13**, 146–165 (2004).
12. F. Melgani, G. Moser, and S. B. Serpico, "Unsupervised change detection methods for remote sensing images," *Opt. Eng.* **41**, 3288–3297 (2002).
13. L. Hansen and P. Salamon, "Neural networks ensembles," *IEEE Trans. Pattern Anal. Mach. Intell.* **12**, 993–1001 (1990).
14. T. K. Ho, J. J. Hull, and S. N. Srihari, "Decision combination in multiple classifier systems," *IEEE Trans. Pattern Anal. Mach. Intell.* **16**, 66–75 (1994).
15. J. A. Benediktsson, J. R. Sveinsson, and P. H. Swain, "Hybrid consensus theoretic classification," *IEEE Trans. Geosci. Remote Sens.* **GE–35**, 833–843 (1997).
16. K. Woods, K. Bowyer, and W. P. Kegelmeyer, "Combination of multiple classifiers using local accuracy estimates," *IEEE Trans. Pattern Anal. Mach. Intell.* **19**, 405–410 (1997).
17. J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas, "On combining classifiers," *IEEE Trans. Pattern Anal. Mach. Intell.* **20**, 226–239 (1998).
18. L. I. Kuncheva, "A theoretical study on six classifier fusion strategies," *IEEE Trans. Pattern Anal. Mach. Intell.* **24**, 281–286 (2002).
19. Y. Freund and R. E. Shapire, "Experiments with a new boosting algorithm," in *Proc. 13th Int. Conf. Machine Learning* (1996).
20. R. E. Shapire, Y. Freund, P. Bartlett, and W. S. Lee, "Boosting the margin: a new explanation for the effectiveness of voting methods," *Proc. 14th Int. Conf. Machine Learning* (1997).
21. L. Breiman, "Bagging predictors," Technical Report 421, Dept. of Statistics, Univ. of California at Berkeley (1994).
22. L. I. Kuncheva, "Fuzzy versus nonfuzzy in combining classifiers designed by boosting," *IEEE Trans. Fuzzy Syst.* **11**, 729–741 (2003).
23. D.-S. Lee and S. N. Srihari, "Handprinted digit recognition: a comparison of algorithms," in *Proc. 3rd Int. Workshop Frontiers Handwriting Recognition*, pp. 153–162, Buffalo, NY (1993).
24. S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," *IEEE Trans. Pattern Anal. Mach. Intell.* **6**, 721–741 (1984).
25. G. R. Cross and A. K. Jain, "Markov random field texture models," *IEEE Trans. Pattern Anal. Mach. Intell.* **5**, 25–39 (1983).
26. S. Krishnamachari and R. Chellappa, "Multiresolution Gauss-Markov random field models for texture segmentation," *IEEE Trans. Image Process.* **6**, 251–267 (1997).
27. Z. Kato, J. Zerubia, and M. Berthod, "Unsupervised parallel image classification using markovian models," *Pattern Recogn.* **32**, 591–604 (1999).
28. A. H. S. Solberg, T. Taxt, and A. K. Jain, "A Markov random field model for classification of multisource satellite imagery," *IEEE Trans. Geosci. Remote Sens.* **GE–34**, 100–113 (1996).
29. F. Melgani and S. B. Serpico, "A Markov random field approach to spatio-temporal contextual image classification," *IEEE Trans. Geosci. Remote Sens.* **GE–41**, 2478–2487 (2003).
30. R. C. Dubes and A. K. Jain, "Random field models in image analysis," *J. Appl. Stat.* **16**, 131–163 (1989).
31. V. Kolmogorov and R. Zabih, "What energy functions can be minimized via graph cuts?," *IEEE Trans. Pattern Anal. Mach. Intell.* **26**, 147–159 (2004).



Farid Melgani received his State Engineer degree in electronics from the University of Batna, Algeria, in 1994, his MSc degree in electrical engineering from the University of Baghdad, Iraq, in 1999, and his PhD degree in electronic and computer engineering from the University of Genoa, Italy, in 2003. From 1999 to 2002, he cooperated with the Signal Processing and Telecommunications Group, Department of Biophysical and Electronic Engineering, University of Genoa. He is currently an assistant professor of telecommunications at the University of Trento, Italy, where he teaches pattern recognition, radar remote-sensing systems, and digital transmission. His research interests are processing and pattern recognition techniques applied to remote sensing and biomedical images (classification, regression, multitemporal analysis, and data fusion). He is coauthor of more than 50 scientific publications. He is a referee for several international journals and has served on the scientific committees of several international conferences. He is an associate editor of the *IEEE Geoscience and Remote Sensing Letters*.