DOI 10.1007/s11263-008-0153-5

On Local Region Models and a Statistical Interpretation of the Piecewise Smooth Mumford-Shah Functional

Thomas Brox · Daniel Cremers

Received: 5 September 2007 / Accepted: 25 June 2008 / Published online: 19 July 2008

Abstract The Mumford-Shah functional is a general and quite popular variational model for image segmentation. In particular, it provides the possibility to represent regions by smooth approximations. In this paper, we derive a statistical interpretation of the full (piecewise smooth) Mumford-Shah functional by relating it to recent works on local region statistics. Moreover, we show that this statistical interpretation comes along with several implications. Firstly, one can derive extended versions of the Mumford-Shah functional including more general distribution models. Secondly, it leads to faster implementations. Finally, thanks to the analytical expression of the smooth approximation via Gaussian convolution, the coordinate descent can be replaced by a true gradient descent.

Keywords Segmentation · Variational methods · Statistical methods · Regularization

1 Introduction

Image segmentation has been one of the most studied problems in image analysis research. Having been handled in a rather heuristic manner for a long time, three seminal works initiated a more systematic approach to the problem: the Bayesian formulation of Geman and Geman (1984), and the two variational formulations in Kass et al. (1988) and Mumford and Shah (1989). All these works replaced the formerly purely algorithmic description of a segmentation technique by the formulation as an optimization problem. This systematic description based on sound mathematical concepts has considerably improved the understanding of image segmentation and supported the development of many new models and algorithms.

Especially in case of the Mumford-Shah functional there has initially been a large gap between its sound theoretical formulation and efficient ways to find minimizers in practice. Although Mumford and Shah (1989) comprises almost 100 pages, there is not a single suggestion on how to implement the underlying segmentation framework. This large gap between theory and practice has been bridged by the works of Ambrosio and Tortorelli (1990), Morel and Solimini (1994), as well as the use of level set representations of contours by Caselles et al. (1993), Chan and Vese (2001), and Paragios and Deriche (2002). Most of the works building upon the Mumford-Shah functional are based on a simplified version, the so-called *cartoon limit*. In this cartoon limit, the image is approximated by a piecewise constant function, contrary to the full Mumford-Shah functional, which aims at a piecewise *smooth* approximation. The cartoon limit has independently been proposed also by Blake and Zisserman (1987), spatially discrete approximations can be minimized by graph cut methods (Greig et al. 1989; Boykov et al. 2001).

Whereas the three above-mentioned approaches to image segmentation are all based on energy minimization, their motivation is quite different. Zhu and Yuille (1996) outlined many relations between these functionals and algorithmic implementations such as region merging or region

We thank the anonymous reviewers for comments leading to improvements of the manuscript, and we gratefully acknowledge funding by the German Research Foundation (DFG) under the project Cr250/1.

T. Brox (✉) · D. Cremers
Computer Vision Group, University of Bonn, Römerstr. 164,
53117 Bonn, Germany
e-mail: brox@cs.uni-bonn.de

D. Cremers
e-mail: dcremers@cs.uni-bonn.de

growing. In particular, they established a link between the statistical maximum a-posteriori approach by Geman and Geman and the cartoon limit of the Mumford-Shah functional. Based on the statistical motivation of the maximum a-posteriori approach, Zhu and Yuille suggested a generalization of the cartoon model, where the constant approximation of image regions is replaced by arbitrary intensity distributions. This formulation was used particularly in level set based segmentation approaches where full Gaussian distributions (Rousson and Deriche 2002), Laplace distributions (Heiler and Schnörr 2005), and nonparametric kernel densities (Kim et al. 2005; Cremers and Rousson 2007) have been suggested. For a recent review of statistical approaches to integrate color, texture, motion and shape, we refer to Cremers et al. (2007).

Zhu and Yuille established relations between statistical methods and the cartoon limit of the Mumford-Shah functional, yet in their work, they ignored the part of the functional that allows also for piecewise *smooth* approximations. In the present paper, we complete their work by showing that the Mumford-Shah functional can be interpreted as a first-order approximation of a specific maximum a-posteriori model, where pixel intensities are not, as usual, identically distributed but where the distribution varies with the position in the image. Such local region statistics have recently been introduced in the scope of medical image segmentation (Taron et al. 2004) and silhouette based 3D tracking (Brox et al. 2005). The statistical interpretation we derive is considerably different from the one in Tsai et al. (2001). Whereas Tsai et al. (2001) focus on the joint segmentation and denoising task, we consider here pure segmentation, where the smooth approximation is a latent variable. In contrast to the interpretation in Tsai et al. (2001), the new one has several practical implications in the scope of segmentation. Firstly, it allows to generalize the Mumford-Shah functional. Such generalizations can be derived from various local statistical models by using the equivalence in the opposite direction. In particular, we propose a functional that approximates the input intensity by a piecewise smooth Gaussian distribution including mean *and* variance. Secondly, one gets access to numerical implementations that are much more efficient than the usual way of solving a large linear system for each region. Our comparison of five implementations reveals significant speedups. Finally, the analytical expression for the smooth approximation given the region contours in the statistical model allows for a gradient descent in the contour as opposed to a coordinate descent in the contour and the regional intensity approximation. Since such an analytical expression is lacking for the Mumford-Shah functional, a gradient descent cannot be derived without the relationship to local region statistics.

We previously presented our statistical interpretation of the Mumford-Shah functional at a conference (Brox and Cremers 2007b). In the present paper we extend this work by focusing on the practical implications of this interpretation.

The remainder of this paper is organized as follows. In Sects. 2 and 3 we briefly review the Mumford-Shah functional and the statistical approach to image segmentation via maximum a-posteriori estimation of contours, respectively. In Sect. 4, we then show an equivalence between these two approaches. From this we obtain a new statistical interpretation of the Mumford-Shah functional that allows extending the functional and to employ more efficient implementations. These extensions are described in Sect. 5. Moreover, we derive the Euler-Lagrange equations of local region statistics and compare the gradient descent to the usual coordinate descent. Supplementary online material (Brox and Cremers 2007a) further contains a brief demonstration of local region statistics in the scope of contour tracking. The paper concludes with a summary in Sect. 6.

2 The Mumford-Shah Functional

The idea of Mumford and Shah was to combine image denoising and segmentation by a functional that simultaneously seeks a piecewise smooth approximation $u : (\Omega \subset \mathbb{R}^2) \to \mathbb{R}$ of the image $I : (\Omega \subset \mathbb{R}^2) \to \mathbb{R}$ and a minimal edge set K that separates the non-smooth parts from each other. This can be expressed as minimization of the functional

$$E(u, K) = \int_\Omega (u - I)^2 \mathbf{dx} + \lambda \int_{\Omega - K} |\nabla u|^2 \mathbf{dx} + \nu |K|, \qquad (1)$$

where $\lambda \geq 0$ and $\nu \geq 0$ are constant weighting parameters. Since our focus lies on image segmentation, we will only consider edge sets which are sets of rectifiable closed curves (Morel and Solimini 1994). In this case, the edge set partitions the image into an a priori unspecified number of disjoint regions Ω_i, with $\Omega = \bigcup_i \Omega_i$, each being approximated by a smooth function $u_i : \Omega_i \to \mathbb{R}$.

An interesting special case arises for $\lambda \to \infty$, where u is required to be piecewise constant. This case, already discussed in Mumford and Shah (1989), is known as the *cartoon limit* and can be written in short form

$$E(u, K) = \sum_i \int_{\Omega_i} (u_i - I)^2 \mathbf{dx} + \nu_0 |K|, \qquad (2)$$

where Ω_i denotes the piecewise constant regions separated by K and ν_0 is the rescaled version of the parameter ν in (1). In this limiting case, u_i is no longer a function but collapses to a single value. Due to the quadratic penalizer, given Ω_i, u_i becomes the mean of I within Ω_i. A related approach was independently developed by Blake and Zisserman (1987). In the spatially discrete case, (2) is related to the Potts model (Potts 1952).

The model in (2) can be simplified further by assuming an a priori fixed number of regions N. In particular, the case $N = 2$ and its level set formulation by Chan and Vese (2001) has become very popular. A discrete version of the binary case has been introduced by Lenz and Ising for modeling ferromagnetism already in the 1920s (Lenz 1920; Ising 1925).

3 Maximum A-Posteriori Model and Local Region Statistics

An alternative approach to image segmentation can be derived using Bayes' rule

$$p(K|I) = \frac{p(I|K)p(K)}{p(I)}. \tag{3}$$

Here one seeks for a partitioning by the edge set K that maximizes the a-posteriori probability given the image I. The first factor in the numerator is in general approximated by an intensity distribution in the regions $i = 1, \ldots, N$ separated by K. The second factor is the a-priori probability of a certain partitioning K. Usually, the total length of the edge set K is assumed to be small,

$$p(K) = \exp(-\nu_B|K|), \tag{4}$$

but other more sophisticated shape priors can be integrated here, as well (Cremers et al. 2002). Assuming independence of intensities at different locations $\mathbf{x}$, one can write a continuous product with $\mathbf{dx}$ being the infinitesimal bin size. With the partitioning of Ω by the edge set K into disjoint regions $\Omega = \bigcup_i \Omega_i$, $\Omega_i \cap \Omega_j = \emptyset, \forall i \neq j$, the product can be separated into products over the regions:

$$\begin{aligned} p(I|K) &= \prod_{\mathbf{x}\in\Omega} p(I(\mathbf{x})|K,\mathbf{x})^{\mathbf{dx}} \\ &= \prod_i \prod_{\mathbf{x}\in\Omega_i} p(I(\mathbf{x})|\mathbf{x}, \mathbf{x}\in\Omega_i)^{\mathbf{dx}}. \end{aligned} \tag{5}$$

For convenience we define the conditional probability density to encounter an intensity s at position $\mathbf{x}$ given that $\mathbf{x} \in \Omega_i$ as

$$p_i(s,\mathbf{x}) := p(s|\mathbf{x}, \mathbf{x}\in\Omega_i). \tag{6}$$

Note that we have here a family of probability densities $p_i(s,\mathbf{x})$ for all $\mathbf{x} \in \Omega$, i.e.,

$$\begin{aligned} &p_i(s,\mathbf{x}) : \mathbb{R} \to \mathbb{R}_0^+, \quad p_i(s,\mathbf{x}) \geq 0, \\ &\int_{\mathbb{R}} p_i(s,\mathbf{x})ds = 1, \quad \forall s \in \mathbb{R}, \forall \mathbf{x} \in \Omega. \end{aligned} \tag{7}$$

In general, it is preferable to express the maximization of (3) by the minimization of its negative logarithm. With the above assumptions, this leads to a generalized version of the cartoon limit (Zhu and Yuille 1996):

$$E(K) = \sum_i \int_{\Omega_i} -\log p_i(I(\mathbf{x}),\mathbf{x})\mathbf{dx} + \nu_B|K|. \tag{8}$$

A typical model for the probability densities p_i is a homogeneous Gaussian distribution in each region Ω_i:

$$p_i(s,\mathbf{x}) \equiv p_i(s) = \frac{1}{\sqrt{2\pi}\sigma_i}\exp\left(-\frac{(s-\mu_i)^2}{2\sigma_i^2}\right), \tag{9}$$

where μ_i and σ_i denote the mean and standard deviation of I in region Ω_i. Alternatively, a Laplace distribution (Heiler and Schnörr 2005) and a nonparametric density (Kim et al. 2005) have been suggested. All these models apply the same probability density to all points in a region. Hence, we will call them spatially *homogeneous* region models.

In contrast, *local* region models drop the assumption of identically distributed pixel intensities within a region.[1] Taking the spatial position into account, there is in general a different probability density at each point $\mathbf{x}$ in the region, i.e., $\mu_i, \sigma_i \in \mathbb{R}$ turn into functions $\mu_i, \sigma_i : \Omega_i \to \mathbb{R}$. Estimation of these functions can be achieved using a window function, e.g. a Gaussian G_ρ with standard deviation ρ, that restricts the estimation to points within this window (Brox 2005):

$$\begin{aligned} \mu_i(\mathbf{x}) &= \frac{\int_{\Omega_i} G_\rho(\mathbf{x}-\zeta)I(\zeta)\,d\zeta}{\int_{\Omega_i} G_\rho(\mathbf{x}-\zeta)\,d\zeta}, \\ \sigma_i(\mathbf{x}) &= \frac{\int_{\Omega_i} G_\rho(\mathbf{x}-\zeta)(I(\zeta)-\mu_i(\mathbf{x}))^2\,d\zeta}{\int_{\Omega_i} G_\rho(\mathbf{x}-\zeta)\,d\zeta}. \end{aligned} \tag{10}$$

Obviously, the local region model converges to the corresponding homogeneous model for $\rho \to \infty$.

Rather than applying a parametric model for $p_i(s,\mathbf{x})$, one can also set up a local nonparametric model via a kernel density estimator (Brox 2005). Estimating the densities with the Parzen method (Parzen 1962) from samples in the local neighborhood G_ρ, yields

$$p_i(s,\mathbf{x}) = \frac{\int_{\Omega_i} G_\rho(\mathbf{x}-\zeta)K_h(s-I(\zeta))\,d\zeta}{\int_{\Omega_i} G_\rho(\mathbf{x}-\zeta)\,d\zeta}, \tag{11}$$

where K_h is a suitable kernel function of width h. Often K_h is chosen to be the Gaussian kernel with standard deviation h. Such a local, nonparametric region model yields a very general descriptor of regions.

[1] This should not be confused with dropping the independence assumption stated above.

4 Statistical Interpretation of the Mumford-Shah Functional

The maximum a-posteriori model from the last section is quite flexible in the choice of the probability density function. It further yields a nice statistical interpretation of the model assumptions and allows for the sound integration of a-priori information. The reader may have noticed similarities between the models in Sects. 2 and 3. Indeed there has been proven an equivalence between the cartoon model and the MAP estimate. Having a fixed standard deviation $\sigma = \sqrt{0.5}$ for all regions and setting $\nu_B = \nu_0$, the MAP energy from (8) becomes exactly the cartoon model in (2) (Zhu and Yuille 1996).

With this equivalence in mind, is there a choice of the probability density function that relates the Bayesian model to the full, piecewise smooth Mumford-Shah functional stated in (1)? A straightforward statistical interpretation has been given in Tsai et al. (2001), where the Bayesian model is extended to $p(K, u|I) \propto p(I|K, u)p(u|K)p(K)$. In this interpretation, both the edge set K and the smooth approximation u are sought and the additional term in the full Mumford-Shah functional can be regarded as a conditional prior on u. While this interpretation is fully satisfactory for the joint segmentation and denoising problem, the prior $p(u|K)$ is of little use, if the pure segmentation problem (3) is considered, where u is only an auxiliary variable. In the following, we give a different statistical interpretation than in Tsai et al. (2001) that neglects the denoising issue and focuses on segmentation. In Sect. 5 it will turn out that this interpretation has many practical implications that in their majority cannot be derived from the interpretation in Tsai et al. (2001).

The new interpretation is based on the fact that (1) explicitly allows the approximation u to vary within a region. Clearly, homogeneous region statistics cannot model this aspect, but local region statistics do. Hence, having in mind that the equivalence of the Bayesian model and the cartoon model was established for a homogeneous Gaussian region model with fixed standard deviation, we take a closer look at the *local* Gaussian model, again with fixed standard deviation. The decisive observation is that the local mean in (10) is a convolution of the image I with the Gaussian function G_ρ including a normalizing denominator for the case that the window hits the boundary of Ω_i. This normalization only ensures preservation of the average gray value of μ_i in the domain Ω_i.

In order to relate this model to the Mumford-Shah functional, we will formulate the filtering operation in a regularization framework. Yuille and Grzywacz (1988) as well as Nielsen et al. (1997) showed that the outcomes of some linear filters are exact minimizers of certain energy functionals with an infinite sum of penalizer terms of arbitrarily high order. More precisely, it was shown in Nielsen et al. (1997) that filtering an image I with the filter

$$\hat{h}(\omega) = \frac{1}{1 + \sum_{k=1}^{\infty} \alpha_k \omega^{2k}} \tag{12}$$

given in the frequency domain, yields the minimizer of the following energy functional:

$$E(u) = \int_{\mathbb{R}} \left((u - I)^2 + \sum_{k=1}^{\infty} \alpha_k \left(\frac{d^k u}{dx^k} \right)^2 \right) dx. \tag{13}$$

In particular, this includes for $\alpha_k = \frac{\lambda^k}{k!}$, the Gaussian filter

$$\hat{h}(\omega, \lambda) = \frac{1}{1 + \sum_{k=1}^{\infty} \frac{\lambda^k}{k!} \omega^{2k}} = \exp(-\lambda \omega^2). \tag{14}$$

This filter corresponds to the Gaussian G_ρ with standard deviation $\rho = \sqrt{2\lambda}$ in the spatial domain. Nielsen et al. (1994) further showed that for Cartesian invariants, such as the Gaussian, this correspondence can be generalized to higher dimensions. Therefore, the convolution result in (10) is the exact minimizer of

$$E(\mu_i) = \int_{\Omega_i} \left((\mu_i - I)^2 + \sum_{k=1}^{\infty} \frac{\lambda^k}{k!} \sum_{j_1 + j_2 = k} \left(\frac{d^k \mu_i}{dx^{j_1} dy^{j_2}} \right)^2 \right) \mathbf{dx} \tag{15}$$

with natural boundary conditions.

Based on these findings, we can generalize the piecewise constant case. We plug the local Gaussian probability density with fixed standard deviation $\sigma = \sqrt{0.5}$ into the Bayesian model in (8):

$$\begin{aligned} E_B(K) &= \sum_i \int_{\Omega_i} \frac{1}{2} \log(2\pi\sigma^2) + \frac{(I(\mathbf{x}) - \mu_i(\mathbf{x}))^2}{2\sigma^2} \mathbf{dx} + \nu_B |K| \\ &= \sum_i \int_{\Omega_i} (I(\mathbf{x}) - \mu_i(\mathbf{x}))^2 \mathbf{dx} + \nu_B |K| + \text{const.} \end{aligned} \tag{16}$$

The means μ_i have been defined in (10) as the results of local convolutions. As we have just found, this convolution result is the minimizer of (15). Hence, we can write the Bayesian energy as:

$$\begin{aligned} E_B(\mu, K) = \sum_i \int_{\Omega_i} \Bigg(& (\mu_i - I)^2 \\ & + \sum_{k=1}^{\infty} \frac{\lambda^k}{k!} \sum_{j_1 + j_2 = k} \left(\frac{d^k \mu_i}{dx^{j_1} dy^{j_2}} \right)^2 \Bigg) \mathbf{dx} \\ & + \nu_B |K|. \end{aligned} \tag{17}$$

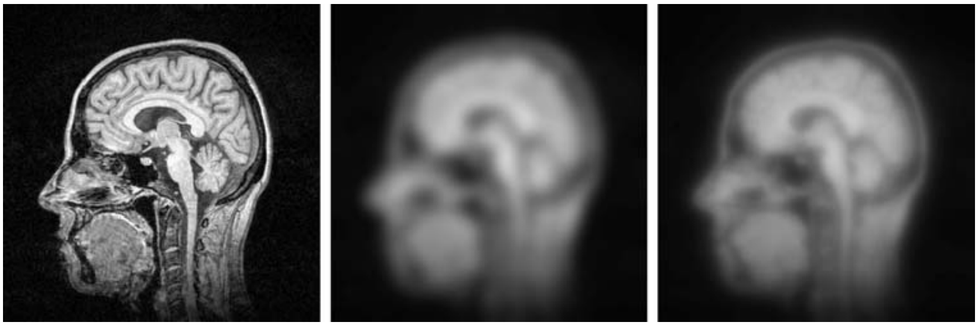

Fig. 1 Comparison of regularization with and without higher order penalizers. *Left*: Original image. *Center*: Smoothing result with the regularizer in (17) (Gaussian smoothing) for $\lambda = 20$. *Right*: Smoothing results with the regularizer in (18) for $\lambda = 20$

Neglecting all penalizer terms of order $k > 1$ yields

$$E_{MS}(\mu, K) = \sum_i \int_{\Omega_i} ((\mu_i - I)^2 + \lambda|\nabla\mu_i|^2)\,\mathbf{dx} + \nu_B|K| + \text{const.} \tag{18}$$

which states exactly the Mumford-Shah functional in (1). Consequently, *minimizing the full piecewise smooth Mumford-Shah functional is equivalent to a first-order approximation of a Bayesian a-posteriori maximization based on local region statistics*. In particular, it is the approximation of the Bayesian setting with a Gaussian distribution, fixed standard deviation $\sigma = \sqrt{0.5}$, and a Gaussian windowing function where $\rho = \sqrt{2\lambda}$ and $\nu_B = \nu$.

The main effect of neglecting the higher order terms is that the minimizers μ_i of the functional in (18) are less smooth than those of the functional in (17). Figure 1 depicts a comparison in case of the whole image domain being a single region. Obviously, the visual difference is almost negligible, and it can be further reduced by choosing λ in the first-order approximation slightly larger than in the regularizer containing the infinite sum of penalizers.

5 Consequences of the Statistical Interpretation

5.1 The Mumford-Shah Functional Including Variance

The statistical interpretation of the Mumford-Shah functional places us in a position to modify the original functional in a way that it can deal with more general distributions. For instance, one may cast doubt on the inherent assumption of having a fixed variance in the whole image. In the statistical formulation, taking the variance into account is very easy, as shown in Sect. 3. Hence, we can take the statistical model and express the convolutions by regularization formulations in order to obtain a corresponding extended Mumford-Shah functional.

With the full Gaussian model, the probability densities

$$p_i(s, \mathbf{x}) = \frac{1}{\sqrt{2\pi}\sigma_i(\mathbf{x})} \exp\left(\frac{(s - \mu_i(\mathbf{x}))^2}{2\sigma_i(\mathbf{x})^2}\right) \tag{19}$$

depend on two functions $\mu_i(\mathbf{x})$ and $\sigma_i(\mathbf{x})$ given by (10). For $\rho \to \infty$ they are the mean and standard deviation of I in Ω_i, i.e., the minimizers of

$$\int_{\Omega_i} \left(\frac{(\mu_i - I)^2}{2\sigma_i^2} + \frac{1}{2}\log(2\pi\sigma_i^2) + \lambda(|\nabla\mu_i|^2 + |\nabla\sigma_i|^2)\right)\mathbf{dx} \tag{20}$$

for $\lambda \to \infty$. This yields a generalized cartoon model. For $\rho \ll \infty$ we make use of the relation between Gaussian convolution and regularization stated in the previous section and obtain $\mu_i(\mathbf{x})$ and $\sigma_i(\mathbf{x})$ as the minimizers of

$$\begin{aligned} E(\mu_i, \sigma_i) = &\int_{\Omega_i} \left(\frac{(\mu_i - I)^2}{2\sigma_i^2} + \frac{1}{2}\log(2\pi\sigma_i^2)\right)\mathbf{dx} \\ &+ \int_{\Omega_i} \sum_{k=1}^{\infty} \frac{\lambda^k}{k!} \sum_{j_1+j_2=k} \left(\frac{d^k\mu_i}{dx^{j_1}dy^{j_2}}\right)^2 \mathbf{dx} \\ &+ \int_{\Omega_i} \sum_{k=1}^{\infty} \frac{\lambda^k}{k!} \sum_{j_1+j_2=k} \left(\frac{d^k\sigma_i}{dx^{j_1}dy^{j_2}}\right)^2 \mathbf{dx}. \end{aligned} \tag{21}$$

Based on the observation in Sect. 4, a qualitatively similar approach is obtained by neglecting the penalizer terms with $k > 1$, which yields an extended version of the Mumford-Shah functional:

$$\begin{aligned} E(\mu, \sigma, K) = &\int_{\Omega} \left(\frac{(\mu - I)^2}{2\sigma^2} + \frac{1}{2}\log(2\pi\sigma^2)\right)\mathbf{dx} \\ &+ \lambda \int_{\Omega - K} (|\nabla\mu|^2 + |\nabla\sigma|^2)\mathbf{dx} + \nu|K|. \end{aligned} \tag{22}$$

One immediate advantage of this functional versus the original from Mumford and Shah (1989) is the possibility to distinguish regions being equal in their mean value but differing in their variance. Such an example is shown in Fig. 2. Since the difference in the variance helps driving the contour, local minimization of the extended functional succeeds in splitting the image correctly. The original Mumford-Shah functional missing this support, on the other hand, results

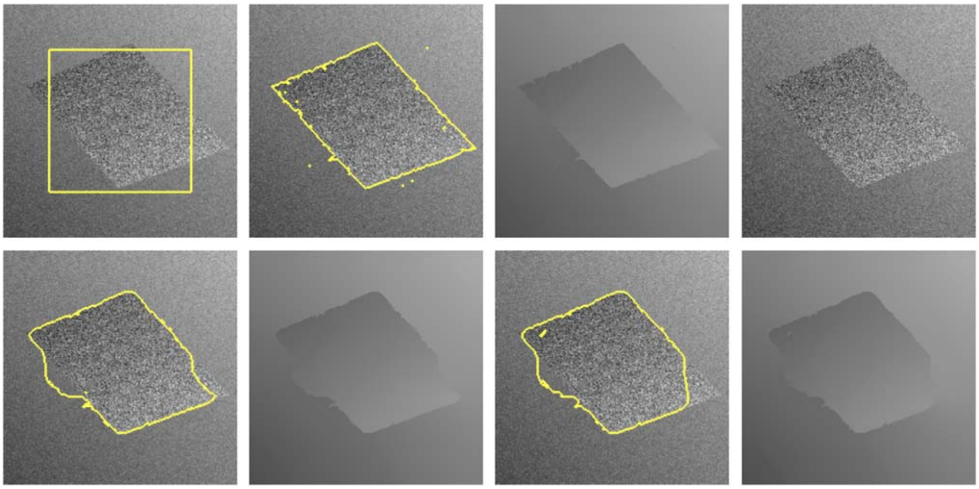

Fig. 2 *Top row*, from *left* to *right*: (**a**) Original image of size 200 × 200 pixels with the initial contour. (**b**) Contour obtained with the extended Mumford-Shah functional in (22) modeling the variance ($\lambda = 72$, $\nu = 2$). (**c**) Approximated mean μ. (**d**) Image generated by sampling from the approximated mean and variance. *Bottom row*, from *left* to *right*: For comparison, results of the traditional Mumford-Shah functional (no variance included). (**e**) Contour for $\nu = 2900$. (**f**) Approximated mean for $\nu = 2900$. (**g**) Contour for $\nu = 3100$. (**h**) Approximated mean for $\nu = 3100$. There is no choice of ν that perfectly captures the region

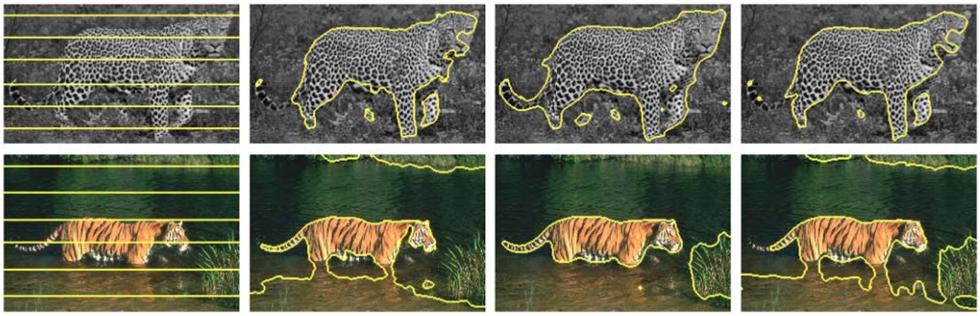

Fig. 3 Two texture segmentation examples. From *left* to *right*: (**a**) Original images of size 256 × 151 and 241 × 161 pixels, respectively, with the initial contour. (**b**) Result with the traditional Mumford-Shah functional (no variance included). (**c**) Result with the functional in (23) modeling the variance. (**d**) Result with the nonparametric model in (24), a bin width of 4 and a Parzen kernel width of 1. The implicit weighting in the functional (23) yields favorable segmentations. Although the nonparametric model includes an (even more sophisticated) weighting, local minima yield visually less pleasant segmentations

in a partially unpleasant solution. Also note the good reconstruction of the original image from the estimated mean and variance function shown in Fig. 2(d). This illustrates the larger descriptive power of the extended functional. Results have been obtained with a level set implementation and a restriction of the model to two regions.

Including the variance becomes even more interesting in the vector-valued case

$$\begin{aligned} E(\mu,\sigma,K) = \sum_{k=1}^{M}\int_{\Omega}\left(\frac{(\mu_k - I_k)^2}{2\sigma_k^2} + \frac{1}{2}\log(2\pi\sigma_k^2)\right)\mathbf{dx} \\ + \lambda\sum_{k=1}^{M}\int_{\Omega-K}(|\nabla\mu_k|^2 + |\nabla\sigma_k|^2)\mathbf{dx} \\ + \nu|K|, \end{aligned} \tag{23}$$

where $I : \Omega \to \mathbb{R}^M$, $\mu : \Omega \to \mathbb{R}^M$, and $\sigma : \Omega \to \mathbb{R}^M$. In this case, the separate variance functions estimated for each channel act as implicit weights. Consequently, the influence of each channel only depends on its discriminative properties and not on the magnitude of the channel values. This allows for the sound integration of different input channels with different contrast and noise levels.

Figure 3 demonstrates this property by comparing the outcome of texture segmentation for the functional in (23) and the same functional with the standard deviation set fixed. Four texture feature channels according to Brox and Weickert (2006) have been supplemented to the gray value and color channels, respectively. In order to attenuate the severe influence of local minima in the piecewise smooth Mumford-Shah functional, a coarse-to-fine strategy was ap-

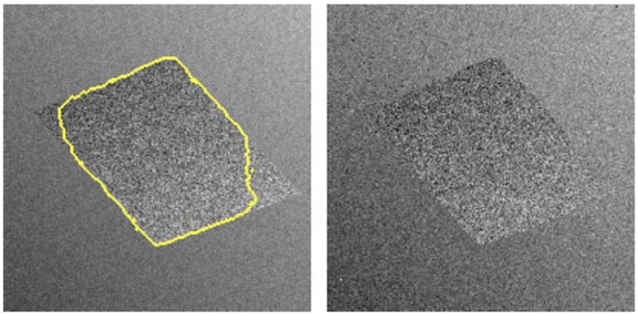

Fig. 4 Same experiment as in Fig. 2 but with the nonparametric model from (24). The bin width for the Parzen estimator was 4 and the kernel was a Gaussian with standard deviation 1. $\lambda = 72$ and $\nu = 4$. The approximation of the input image obtained by sampling from the estimated densities is quite good. The contour, however, does not perfectly separate the two regions due to local minima in the contour evolution

plied. In particular, starting from the cartoon limit, the parameter λ was slowly decreased until it reached the final value of $\lambda = 72$. Figure 3 shows that including the variance, thereby including implicit channel weights, yields favorable segmentations.

5.2 A Nonparametric Mumford-Shah Functional

The Mumford-Shah functional can be further generalized by introducing nonparametric density models. Let p be family of probability density functions as defined in (7). These general densities can be described by a kernel density estimator leading to a nonparametric model similar to the one in (11). The following functional involving the kernel K_h prefers segmentations such that p is spatially smooth in each region:

$$\begin{aligned} E(K) = \sum_i \Bigg(\int_{\Omega_i} & -\log p(I(\mathbf{x}), \mathbf{x})\, \mathbf{dx} \\ & + \int_{\mathbb{R}} \int_{\Omega_i} (p(s, \mathbf{x}) - K_h(s - I(\mathbf{x})))^2 \\ & + \lambda |\nabla p(s, \mathbf{x})|^2 \, \mathbf{dx} ds \Bigg) + \nu |K|. \end{aligned} \tag{24}$$

Although it is on the first glance not clear whether the property $\int p(s, \mathbf{x}) ds = 1, \forall \mathbf{x}$ holds for the densities involved here, one can prove this by regarding gradient descent for enforcing the smoothness. Let $p_0(s, \mathbf{x}) = K_h(s - I(\mathbf{x}))$. For all times t in the gradient descent and for all λ, p remains a family of densities because: $\partial_t \int p(s, \mathbf{x}) ds = \int \partial_t p(s, \mathbf{x}) ds = \int \Delta p(s, \mathbf{x}) - (p(s, \mathbf{x}) - p_0(s, \mathbf{x}))/\lambda ds = \Delta \int p(s, \mathbf{x}) ds - \int (p(s, \mathbf{x}) - p_0(s, \mathbf{x}))/\lambda ds = 0$.

Applying a kernel density estimator certainly leads to an extremely general segmentation model. In combination with local optimization schemes this usually yields the drawback of having more local optima in the objective function than simpler models. Figure 4 shows the result of the same experiment as in Fig. 2 but with the nonparametric model. It indicates the problems that may arise from a too general model. Nonetheless, a local nonparametric model can be interesting in case of other optimization techniques or in case of λ being rather large combined with close initializations.

5.3 Efficient Implementation

Given the striking similarity between the first-order regularization and Gaussian convolution (see Fig. 1) and the theorem by Nielsen et al., we can consider to exchange the implementation of regularization and Gaussian convolution. As we show in this section, this can lead to significant speedups. We compared the following implementations for computing the smooth approximation μ_i of each region Ω_i.

Regularization The typical implementation is to minimize

$$\begin{aligned} & E(\mu_i(\mathbf{x})) \\ & \quad = \int_{\Omega_i} \left((\mu_i(\mathbf{x}) - I(\mathbf{x}))^2 \, \mathbf{dx} + \lambda \int_{\Omega} |\nabla \mu_i(\mathbf{x})|^2 \right) \mathbf{dx} \end{aligned} \tag{25}$$

by solving the corresponding Euler-Lagrange equation

$$\mathbf{1}_{\mathbf{x} \in \Omega_i} \cdot (\mu_i(\mathbf{x}) - I(\mathbf{x})) - \lambda \Delta \mu_i(\mathbf{x}) = 0, \tag{26}$$

where Δ denotes the Laplace operator and $\mathbf{1}_{\mathbf{x} \in \Omega_i} = 1$ if $\mathbf{x} \in \Omega_i$ and 0 otherwise. Discretization of (26) yields a linear system. We solved this linear system once by gradient descent (2000 iterations) and once by a unidirectional multigrid scheme employing 25 iterations of successive overrelaxation (SOR) at each grid level. A slightly faster implementation could be achieved with a full multigrid solver.

Gaussian Convolution According to Nielsen et al. (1997), the above regularization is a first-order approximation of Gaussian convolution. Hence, the minimizer in (25) can be well approximated by the normalized convolution

$$\mu_i(\mathbf{x}) = \frac{\int_{\Omega_i} G_\rho(\mathbf{x} - \zeta) I(\zeta)\, d\zeta}{\int_{\Omega_i} G_\rho(\mathbf{x} - \zeta)\, d\zeta}. \tag{27}$$

The naive implementation includes two convolutions with a sampled Gaussian: one convolution of the image and one of the region indicator function in the denominator. For efficiency reasons, the Gaussian was truncated outside the 2ρ-interval.

Recursive Filtering A much more efficient approximation of the above convolution operation is by recursive filtering (Deriche 1990), particularly if ρ is large. The idea of such filters is to recursively propagate information from one part of the image to another in a single forward-backward sweep.

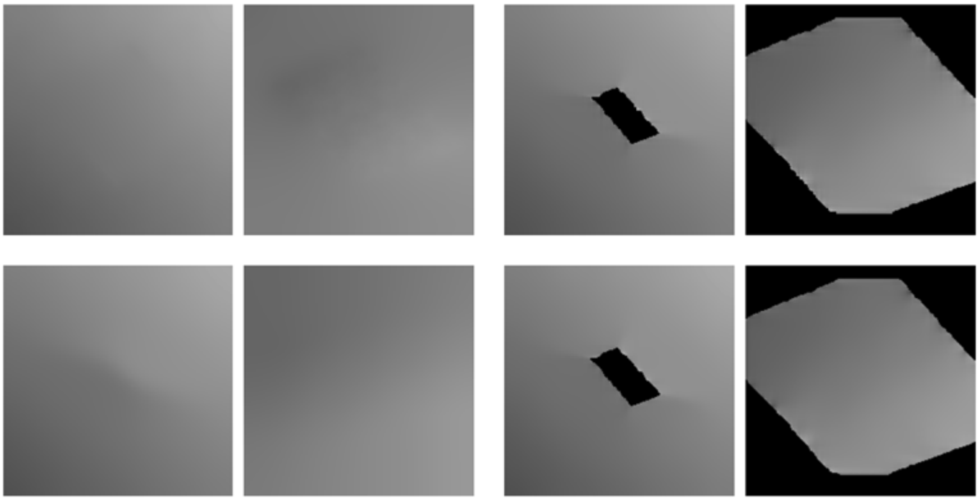

Fig. 5 Smooth region approximations for the image in Fig. 2 obtained with different implementations. *Top left*: Minimization of the regularization functional. *Top right*: Convolution with a Gaussian truncated at 2ρ. *Bottom left*: Recursive filter. *Bottom right*: Four iterations of a box filter. Since the truncated Gaussian and the box filter have only compact support, areas that are far outside the region are not specified

Whereas the Gaussian convolution described above has a time complexity of $O(kN)$, where N is the number of pixels and $k = 4\rho$ the width of the sampled Gaussian, recursive filtering has a complexity of $O(N)$, which is independent of ρ. This is particularly useful in case of the Mumford-Shah functional, where typical values of ρ are in the area of $\rho = 12$ or larger. Recursive filtering in this context has been suggested in Piovano et al. (2007). We implemented the second order recursive filter from Deriche (1990).

Iterated Convolution with a Box Filter Another fast alternative to Gaussian convolution, which has also a time complexity of $O(N)$, is by iterating a box filter, i.e., convolution with the filter mask

$$h(s) = \begin{cases} \frac{1}{2\rho} & \text{if } |s| < \rho, \\ 0 & \text{else.} \end{cases} \tag{28}$$

Infinite convolution of this function with itself yields a Gaussian. However, three iterations are already sufficient to be close to a Gaussian. We applied the filter four times, two times in each direction. Convolution with $h(s)$ can be implemented very efficiently, since

$$\frac{1}{2\rho}\sum_{i=a+1}^{b+1} I(i) = \frac{1}{2\rho}\left(\sum_{i=a}^{b} I(i) + I(b+1) - I(a)\right).$$

Figure 5 depicts the smooth approximations computed for the two regions of the synthetic example in Fig. 2. Clearly, the four implementations yield very similar results, especially inside the respective region. Far outside each region, the computed values start to differ. In particular, as the Gaussian filter and the box filter both have a compact support, values are unspecified for points that are farther than 2ρ from the region. This is irrelevant for local contour evolutions, but global minimization techniques, such as graph cuts (Greig et al. 1989; Boykov et al. 2001) or comparable continuous algorithms (Chan et al. 2006) need values to be specified in the whole image domain.

Table 1 Computation times for computing the smooth approximation and the variance of one region in a 200 × 200 image with $\lambda = 72$, or $\rho = 12$, on a 2 GHz Pentium M processor

Implementation	Computation time	Speedup
Regularization (gradient descent)	2384.04 ms	1
Gaussian convolution	104.74 ms	24
Regularization (cascadic SOR)	68.59 ms	35
Recursive filter	7.40 ms	322
Box filter	7.10 ms	336

Table 1 compares computation times. A naive Gaussian convolution is already 24 times faster than regularization implemented with a simple gradient descent. While a fast implementation with cascadic SOR is faster than naive Gaussian convolution, recursive filtering as well as the iterated box filter provide speedups of more than two orders of magnitude. This clearly demonstrates the advantage of computing the smooth approximations by a convolution expression rather than the regularization framework.

5.4 Exact Shape Gradient

The statistical framework allows for the computation of the gradient with respect of the contour. This becomes possible, since the dependency of the smooth region approximations on the contours are available in an analytic form by means of convolution expressions, in contrast to the Mumford-Shah functional, where these approximations can only be computed numerically. This leads to an exact gradient descent, whereas the usual implementation is only a coordinate descent. Precise shape gradients including secondary terms are not unusual, though they are rarely implemented. In Piovano et al. (2007) the secondary terms for a local Gaussian distribution with fixed variance have been derived and brought in a form that allows for an efficient implementation. The exact shape gradient with respect to the level set function $\Phi : \Omega \to \mathbb{R}$ can be computed by means of the Gâteaux derivative. In particular

$$\left.\frac{dE(\Phi(\mathbf{x}) + \epsilon h(\mathbf{x}))}{d\epsilon}\right|_{\epsilon=0} = 0 \tag{29}$$

must hold for any test function $h(\mathbf{x})$. We focus here on the contribution of a single region with a local Gaussian distribution and neglect the length constraint, which is known to yield a curvature dependent term. More details can be found in Brox and Cremers (2007a). With the following abbreviations

$$F_1(\mathbf{x}) := (K * H(\Phi))(\mathbf{x}) = \int_\Omega K(\mathbf{x}-\mathbf{y}) H(\Phi(\mathbf{y}))\mathbf{dy},$$

$$F_2(\mathbf{x}) := (K * (H(\Phi)I)(\mathbf{x}) = \int_\Omega K(\mathbf{x}-\mathbf{y}) H(\Phi(\mathbf{y}))I(\mathbf{y})\mathbf{dy} \quad \rightarrow \quad \mu(\mathbf{x}) = \frac{F_2(\mathbf{x})}{F_1(\mathbf{x})},$$

$$F_3(\mathbf{x}) := (K * (H(\Phi)I^2)(\mathbf{x}) = \int_\Omega K(\mathbf{x}-\mathbf{y}) H(\Phi(\mathbf{y}))I^2(\mathbf{y})\mathbf{dy} \quad \rightarrow \quad \sigma^2(\mathbf{x}) = \frac{F_3(\mathbf{x})}{F_1(\mathbf{x})} - \mu^2(\mathbf{x}),$$

$$F_4(\mathbf{x}) := \left(\bar{K} * \frac{H(\Phi)((I-\mu)^2 - \sigma^2)}{\sigma^4 F_1}\right)(\mathbf{x}) = \int_\Omega \frac{K(\mathbf{y}-\mathbf{x})H(\Phi(\mathbf{y}))((I(\mathbf{y}) - \mu(\mathbf{y}))^2 - \sigma^2(\mathbf{y}))}{\sigma^4(\mathbf{y})F_1(\mathbf{y})}\mathbf{dy},$$

$$F_5(\mathbf{x}) := \left(\bar{K} * \frac{H(\Phi)(2I\sigma^2 - 2\mu(I-\mu)^2)}{\sigma^4 F_1}\right)(\mathbf{x}) = \int_\Omega \frac{K(\mathbf{y}-\mathbf{x})H(\Phi(\mathbf{y}))(2I(\mathbf{y})\sigma^2(\mathbf{y}) - 2\mu(\mathbf{y})(I(\mathbf{y}) - \mu(\mathbf{y}))^2)}{\sigma^4(\mathbf{y})F_1(\mathbf{y})}\mathbf{dy},$$

$$\begin{aligned}F_6(\mathbf{x}) &:= \left(\bar{K} * \frac{H(\Phi)\big(\sigma^2\big(\frac{F_3}{F_1} - 2I\mu\big) - (I-\mu)^2(\sigma^2 - \mu^2)\big)}{\sigma^4 F_1}\right)(\mathbf{x}) \\ &= \int_\Omega \frac{K(\mathbf{y}-\mathbf{x})H(\Phi(\mathbf{y}))\big(\sigma^2(\mathbf{y})\big(\frac{F_3(\mathbf{y})}{F_1(\mathbf{y})} - 2I(\mathbf{y})\mu(\mathbf{y})\big) - (I(\mathbf{y}) - \mu(\mathbf{y}))^2(\sigma^2(\mathbf{y}) - \mu^2(\mathbf{y}))\big)}{\sigma^4(\mathbf{y})F_1(\mathbf{y})}\mathbf{dy},\end{aligned}$$

where $\bar{K}$ denotes the mirrored kernel K, one obtains

$$\begin{aligned}&H'(\Phi(\mathbf{x}))\bigg(\frac{(I(\mathbf{x}) - \mu(\mathbf{x}))^2}{2\sigma^2(\mathbf{x})} + \log\sigma(\mathbf{x}) \\ &\quad - \frac{1}{2}(I^2(\mathbf{x})F_4(\mathbf{x}) + I(\mathbf{x})F_5(\mathbf{x}) + F_6(\mathbf{x}))\bigg) = 0.\end{aligned} \tag{30}$$

This equation can be implemented rather efficiently using the filtering techniques mentioned in Sect. 5.3. The first two terms are the usual part considered when applying coordinate descent. The remaining terms take into account the dependency of the distribution on Φ. In the same style one can derive the gradient for the local nonparametric region model:

$$\begin{aligned}&-H'(\Phi(\mathbf{x}))\bigg(\log p(I(\mathbf{x}), \mathbf{x}) \\ &\quad + \int_\Omega \frac{K(\mathbf{y}-\mathbf{x})H(\Phi(\mathbf{y}))(K_h(I(\mathbf{y}) - I(\mathbf{x})) - p(I(\mathbf{y}), \mathbf{y}))}{p(I(\mathbf{y}), \mathbf{y})\int_\Omega K(\mathbf{y}-\mathbf{z})H(\Phi(\mathbf{z}))\mathbf{dz}}\mathbf{dy}\bigg) = 0.\end{aligned} \tag{31}$$

More details on the derivation of the shape gradient and experimental comparisons can be found in the supporting online material (Brox and Cremers 2007a).

6 Conclusions

In this paper, we have derived a statistical interpretation of the well-known Mumford-Shah functional, in particular the full version that allows for smooth region approximations. We have shown that the Mumford-Shah functional can be regarded as a first-order approximation of a maximum a-posteriori model where each region is modeled by the mean estimated in a local Gaussian neighborhood. This interpretation allows to define extended versions of the Mumford-Shah functional, for instance, by considering the variance in the estimation process. Moreover, the relation between the regularization involved in the Mumford-Shah functional and Gaussian convolution provides more efficient implementations. By using box filters or recursive filters the initially slow estimation of the smooth approximation in each region can be accelerated up to two orders of magnitude. Additionally, the statistical model, thanks to its analytic region description, allows to replace the coordinate descent by a gradient descent. Segmentation models based on local regions statistics provide a nice intermediate stage between conventional region based segmentation and edge based approaches: the first class of methods is driven by global differences in the intensity distribution, the second one by very local differences in the intensity (the image gradient). With local region statistics, one is not obliged to choose one of these extremes, but can freely select the scale of homogeneity expected in an application.

References

Ambrosio, L., & Tortorelli, V. (1990). Approximation of functionals depending on jumps by elliptic functionals via Γ-convergence. *Communications on Pure and Applied Mathematics*, *XLIII*, 999–1036.

Blake, A., & Zisserman, A. (1987). *Visual Reconstruction*. Cambridge: MIT Press.

Boykov, Y., Veksler, O., & Zabih, R. (2001). Fast approximate energy minimization via graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *23*(11), 1222–1239.

Brox, T. (2005). *From pixels to regions: partial differential equations in image analysis*. Ph.D. thesis, Faculty of Mathematics and Computer Science, Saarland University, Germany.

Brox, T., & Cremers, D. (2007a). On local region models and the statistical interpretation of the piecewise smooth Mumford-Shah functional. Supplementary online material. Available at http://www-cvpr.iai.uni-bonn.de/pub/pub/brox_ijcv08_sup.pdf.

Brox, T., & Cremers, D. (2007b). On the statistical interpretation of the piecewise smooth Mumford-Shah functional. In F. Sgallari, A. Murli, & N. Paragios (Eds.), *LNCS: Vol. 4485. Scale space and variational methods in computer vision* (pp. 203–213). Berlin: Springer.

Brox, T., & Weickert, J. (2006). A TV flow based local scale estimate and its application to texture discrimination. *Journal of Visual Communication and Image Representation*, *17*(5), 1053–1073.

Brox, T., Rosenhahn, B., & Weickert, J. (2005). Three-dimensional shape knowledge for joint image segmentation and pose estimation. In W. Kropatsch, R. Sablatnig, & A. Hanbury (Eds.), *LNCS: Vol. 3663. Pattern recognition* (pp. 109–116). Berlin: Springer.

Caselles, V., Catté, F., Coll, T., & Dibos, F. (1993). A geometric model for active contours in image processing. *Numerische Mathematik*, *66*, 1–31.

Chan, T., & Vese, L. (2001). Active contours without edges. *IEEE Transactions on Image Processing*, *10*(2), 266–277.

Chan, T., Esedoḡlu, S., & Nikolova, M. (2006). Algorithms for finding global minimizers of image segmentation and denoising models. *SIAM Journal on Applied Mathematics*, *66*(5), 1632–1648.

Cremers, D., & Rousson, M. (2007). Efficient kernel density estimation of shape and intensity priors for level set segmentation. In J.S. Suri, & A. Farag (Eds.), Parametric and Geometric Deformable Models: An application in Biomaterials and Medical Imagery. Berlin: Springer.

Cremers, D., Tischhäuser, F., Weickert, J., & Schnörr, C. (2002). Diffusion snakes: introducing statistical shape knowledge into the Mumford-Shah functional. *International Journal of Computer Vision*, *50*(3), 295–313.

Cremers, D., Rousson, M., & Deriche, R. (2007). A review of statistical approaches to level set segmentation: integrating color, texture, motion and shape. *International Journal of Computer Vision*, *72*(2), 195–215.

Deriche, R. (1990). Fast algorithms for low-level vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *12*, 78–87.

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *6*, 721–741.

Greig, D., Porteous, B., & Seheult, A. (1989). Exact maximum a posteriori estimation for binary images. *Journal of the Royal Statistical Society B*, *51*(2), 271–279.

Heiler, M., & Schnörr, C. (2005). Natural image statistics for natural image segmentation. *International Journal of Computer Vision*, *63*(1), 5–19.

Ising, E. (1925). Beitrag zur Theorie des Ferromagnetismus. *Zeitschrift für Physik*, *31*, 253–258.

Kass, M., Witkin, A., & Terzopoulos, D. (1988). Snakes: active contour models. *International Journal of Computer Vision*, *1*, 321–331.

Kim, J., Fisher, J., Yezzi, A., Cetin, M., & Willsky, A. (2005). A nonparametric statistical method for image segmentation using information theory and curve evolution. *IEEE Transactions on Image Processing*, *14*(10), 1486–1502.

Lenz, W. (1920). Beitrag zum Verständnis der magnetischen Erscheinungen in festen Körpern. *Physikalische Zeitschrift*, *21*, 613–615.

Morel, J.-M., & Solimini, S. (1994). *Variational methods in image segmentation*. Basel: Birkhäuser.

Mumford, D., & Shah, J. (1989). Optimal approximations by piecewise smooth functions and associated variational problems. *Communications on Pure and Applied Mathematics*, *42*, 577–685.

Nielsen, M., Florack, L., & Deriche, R. (1994). *Regularization and scale space* (Technical Report 2352). INRIA, Sophia-Antipolis, France.

Nielsen, M., Florack, L., & Deriche, R. (1997). Regularization, scale-space and edge detection filters. *Journal of Mathematical Imaging and Vision*, *7*, 291–307.

Paragios, N., & Deriche, R. (2002). Geodesic active regions: a new paradigm to deal with frame partition problems in computer vision. *Journal of Visual Communication and Image Representation*, *13*(1/2), 249–268.

Parzen, E. (1962). On the estimation of a probability density function and the mode. *Annals of Mathematical Statistics*, *33*, 1065–1076.

Piovano, J., Rousson, M., & Papadopoulo, T. (2007). Efficient segmentation of piecewise smooth images. In F. Sgallari, A. Murli, & N. Paragios (Eds.), *LNCS: Vol. 4485. Scale space and variational methods in computer vision* (pp. 709–720). Berlin: Springer.

Potts, R. (1952). Some generalized order-disorder transformation. *Proceedings of the Cambridge Philosophical Society*, *48*, 106–109.

Rousson, M., & Deriche, R. (2002). A variational framework for active and adaptive segmentation of vector-valued images. In *Proc. IEEE workshop on motion and video computing* (pp. 56–62), Orlando, Florida.

Taron, M., Paragios, N., & Jolly, M.-P. (2004). Border detection on short axis echocardiographic views using an ellipse driven region-based framework. In *LNCS: Vol. 3216. Medical image computing and computer assisted interventions* (pp. 443–450). Berlin: Springer.

Tsai, A., Yezzi, A., & Willsky, A. (2001). Curve evolution implementation of the Mumford-Shah functional for image segmentation, denoising, interpolation, and magnification. *IEEE Transactions on Image Processing*, *10*(8), 1169–1186.

Yuille, A., & Grzywacz, N. M. (1988). A computational theory for the perception of coherent visual motion. *Nature*, *333*, 71–74.

Zhu, S.-C., & Yuille, A. (1996). Region competition: unifying snakes, region growing, and Bayes/MDL for multiband image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *18*(9), 884–900.