2. Cleaning Dirty Pictures

In this chapter we will discuss and illustrate the previously introduced concepts. We will pursue the process how expectations and restrictions are translated to and incorporated into Bayesian models.

We continue with edge preserving smoothing and image denoising from Section 1.1; it is conceptually simple, can be grasped intuitively, and needs no further theory. Moreover, it has direct applications in diverse important fields like segmentation, motion analysis, remote sensing, tomography, X-ray imaging, and other medical imaging techniques, just to mention some. This will be made clear in Chapter 21 by various examples. The idea of smoothing while preserving significant dissimilarities also penetrates disciplines where the features of interest are not intensities. Important examples are texture segmentation, and classification, treated in Part V of this text. The general philosophy behind is simple: interesting things happen where something changes; discontinuous phenomena are important carriers of information.

Let us return to intensity patterns. Real scenes usually are composed of comparably smooth regions. Noise degradation results in roughness at small scale. To reduce the noise contribution and thereby 'restore' an image, one smoothes data in one or the other way. Global smoothing has the unpleasant property to blur really existing contrast. Hence we aim at boundary preserving methods. We will distinguish between edges as local instances of sharp contrast and, at a somewhat higher level, boundaries as organized strings of edges or 'regular' contours across which contrast is high. Edges and boundaries play an important role in our considerations. They are intimately connected to edge preserving smoothing, since smooth parts of an image on different intensity levels automatically are separated by something we call boundaries, and boundaries are modelled as strings of edges. Conversely, contrast boundaries surround regions of smoothness. Hence edge detection or boundary finding is an important part of image analysis. The Bayesian approach allows one to incorporate notions like smoothness or regularity of boundaries too. This aspect will be addressed in Example 2.4.1.

2.1 Boundaries and Their Information Content

Before we turn to our main subject, some 'philosophical' remarks on boundaries are in order. These are important primitive image features; for instance they provide an indication of the extent of objects and hence together with other features are helpful for higher level image processing. As an example, we consider intensity discontinuities.

Intensity boundaries may carry a considerable amount of information, relevant for the observer or for processing tools. This is evident for line drawings, handwriting, or blueprints. It seems that transmission of rational information crucially relies on boundaries (for feelings and emotions this seems to be different as a look at impressionist paintings like TURNER's suggests).

This can to some extent be made precise, as U. DAUB (1995) and V. AURICH and U. DAUB (1996) show. Their approach may roughly be summarized as follows: Take a digital picture and set a good edge detector to work. Store the boundaries. This requires much less storage capacity than the whole original picture. Add as a little bit more information intensities from the few pixels on a very sparse sub-grid of the original pixel grid. The ratio of memory required to store these data to that required to store the whole picture is called *compression rate*. Now construct a new picture from these sparse data in the following way (we omit technical details): interpolate the values on the fine original pixel grid from those on the sparse grid. To preserve contrast, the interpolation method from a fixed pixel s does not 'see' intensities in pixels t beyond the boundary surrounding s. This results in a picture which inside regions surrounded by a boundary is smoother than the original one.

Fig. 2.1. Natural scene and BPC compression, rate 2.94% ([73])

The information content of boundaries (together with the little extra information of undersampled intensities) is quantified in the following way: Fix

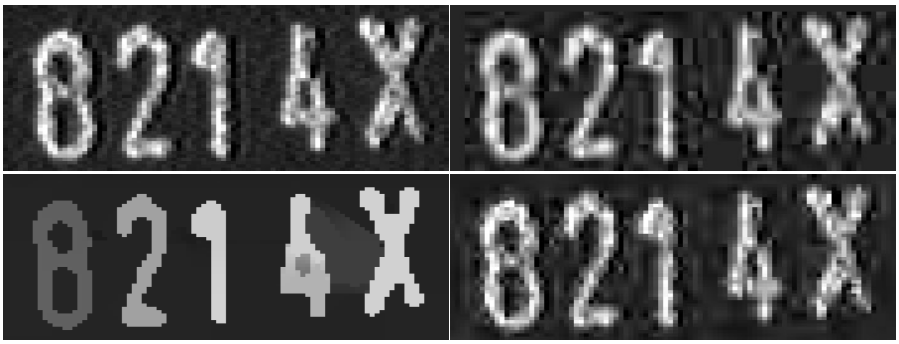


Fig. 2.2. Numerals on a computer chip compressed at rate about 5 %. U.l.: original image; l.l. BPC compression; u.r.: compression with JPEG, based on cosine transform l.r.: EPIC, a wavelet compression; [73]

the boundaries. Store the intensities on more and more sparse sub-grids as long as a human observer accepts the recovered image as a reasonable version of the original one. The resulting compression rate is a measure for the information content of boundaries; if it is 4 % then we threw away 96 % of the original data and still have nearly the same information. Let us abbreviate this method for *boundary preserving compression* by BPC.

Both excellent restoration and compression are impressively illustrated by way of a couple of experiments. In Figure 2.1 a natural scene is processed with BPC. Figure 2.2 compares BPC with JPEG, the former international compression standard (based on cosine transforms), and with the wavelet based method EPIC. Compression based on piecewise smoothing and edge detection is completely different in flavour. Obviously, it is an excellent basis for subsequent processing like number recognition.

2.2 Towards Piecewise Smoothing

Let us now turn to the Bayesian approach to piecewise smoothing of images. This is our first and main example how the Bayesian paradigm can be used in imaging. In this section we develop step by step a prior model for edge preserving smoothing. We shall find simple general principles behind. This is expressed in the following quotation from J.-M. MOREL and S. SOLIMINI (1995), p. XII:

> More than thousand algorithms have been proposed for segmenting images or detecting 'edges'. ... The result of this discussion was unexpected to the authors of this book because they became aware that under the very large diversity of these tools, *there essentially was only one segmentation (or 'edge detection') model.* ... most segmentation algorithms try to minimize ... the same *Segmentation energy.* This

energy measures how smooth the regions are, how faithful the 'analyzed image' to the original image and the obtained 'edges' to the image discontinuities are.

Let us continue from Section 1.1 and start again from the Ising model. Let us first recall the situation for the simplest case: S is a finite square grid, two nodes are declared as neighbours if they are next to each other in one of the four horizontal or vertical directions, and in each node $s \in S$ there is a value g_s equal to ± 1. The Ising energy function

$$K(g) = -\sum_{s \sim t} g_s g_t, \quad g_s = \pm 1, \tag{2.1}$$

measures the excess of unlike over like neighbour pairs. The corresponding Gibbsian prior has the form

$$\Pi(g) = \frac{\exp\left(\sum_{s \sim t} g_s g_t\right)}{\sum_z \exp\left(\sum_{s \sim t} z_s z_t\right)}.$$

Plainly, Π is not affected by addition of a constant to K:

$$\frac{\exp\left(\sum_{s \sim t} g_s g_t\right)}{\sum_z \exp\left(\sum_{s \sim t} z_s z_t\right)} = \frac{\exp\left(\left(\sum_{s \sim t} g_s g_t + C\right)\right)}{\sum_z \exp\left(\left(\sum_{s \sim t} z_s z_t + C\right)\right)}, \ C \in \mathbb{R},$$

and this simple observation allows different interpretations. Continuing with notation from Section 1.1 let $E = \{\{s,t\} \in S \times S : s \sim t\}$. On the one hand, choosing $C = |E|$ we get

$$\begin{aligned} K_L(g) &= -\sum_{s \sim t} g_s g_t - |E| \\ &= -\text{number of like} + \text{number of unlike neighbours} - |E| \\ &= -2 \times \text{number of like neighbour pairs}, \end{aligned} \tag{2.2}$$

This measures the degree of smoothness. On the other hand,

$$\begin{aligned} K_E(g) &= -\sum_{s \sim t} g_s g_t + |E| \\ &= -\text{number of like} + \text{number of unlike neighbours} + |E| \\ &= 2 \times \text{number of unlike neighbour pairs} = 2 \times \text{boundary length}. \end{aligned} \tag{2.3}$$

Hence $K_E(g)$ is proportional to the length of the imaginary boundary between black and white regions. We see that the Ising model comprises two complementary aspects: smoothness of regions and length of boundaries between regions of different intensities. The difference between these aspects becomes more evident if we generalize to multi-valued intensities. Let the g_s take values in a finite set $\mathcal{G}$. Then the straightforward generalization of the boundary approach is the energy function

$$L^\beta(g) = \beta \sum_{s\sim t} \big(1 - \delta(g_s, g_t)\big) = \beta \cdot \text{boundary length}, \quad \beta > 0, \tag{2.4}$$

which again measures the length of the imaginary boundary between regions of different colour. We inserted the parameter β to control the strength of interaction. The MAP estimator tries to keep this boundary length as short as possible. Hence it produces fairly regular boundaries, since wiggly boundaries tend to be long, see Fig. 1.2. The patches of equal colour tend to be large and of regular shape. Note that this can also be interpreted as segmentation of the picture, and in fact, for this model *edge detection = segmentation*. At this point, we are very close to texture segmentation where the g_s are labels associated to certain types of texture. The Gibbs field $\Pi(g) \propto \exp(-L^\beta(g))$ is known as the *Potts model* (R.B. POTTS (1952)).

If $\mathcal{G} \subset \mathbb{R}$, then smoothness may be scored by

$$G^\beta(g) = \beta \sum_{s\sim t} (g_s - g_t)^2, \quad \beta > 0. \tag{2.5}$$

If $g_s = \pm 1$ then $g_s g_t = (g_s - g_t)^2/2 - 1$ and hence (2.5) is compatible with the binary model (2.1). This simultaneously gives a least squares interpretation of (2.1) besides the smoothness interpretation (2.2) and the boundary interpretation (2.3). If one combines the respective priors with a data term, say $D(g, y) = \sum_s (g_s - y_s)^2$, then one arrives at posterior energy functions

$$L^\beta(g, y) = \beta \sum_{s\sim t} \big(1 - \delta(g_s, g_t)\big) + D(g, y) \tag{2.6}$$

$$G^\beta(g, y) = \beta \sum_{s\sim t} (g_s - g_t)^2 + D(g, y). \tag{2.7}$$

The first prior favours sharp boundaries surrounding regions of constant intensity. The second one smoothes gently but blurs contrast. The obvious question is how to draw benefit from both models. If we combine both in one single model then each one should be given a turn where it is superior to the other. To achieve this goal we introduce variables e_{st} which can switch off smoothing and instead allow sharp breaks, and vice versa. For neighbour pixels s, t we introduce the *microedge* $s \sim t$ between s and t. This was illustrated in Fig. 1.4. For each microedge we define 'switch variables'

$$e_{st} = \begin{cases} 0 & \text{no edge between } s \text{ and } t \\ 1 & \text{an edge between } s \text{ and } t \end{cases} \quad \text{for } s \sim t.$$

Generic parameters are now of the form $x = (g, e)$ and the parameter space is $\boldsymbol{X} = \boldsymbol{G} \times \boldsymbol{E}$ with $\boldsymbol{G} = \mathcal{G}^S$ and $\boldsymbol{E} = \{0, 1\}^E$. Now the two models are linked by means of the switches:

$$H(g, e) = \sum_{s\sim t} \Big(\underbrace{\lambda^2 (g_s - g_t)^2}_{\text{smoothing}} \underbrace{(1 - e_{st})}_{\text{on/off}} + \underbrace{\alpha e_{st}}_{\text{penalty}} \Big) + \underbrace{D(g, y)}_{\text{data term}} . \tag{2.8}$$

For the discussion recall from Section 1.1 that low H means 'good'. If $\lambda^2(g_s - g_t)^2 > \alpha$ then it pays off to accept the penalty α for $e_{st} = 1$ and to switch off the smoothing term (i.e. set it to 0). Therefore the penalty for high contrast $|g_s - g_t|$ will not exceed α, and contrast survives if the distance to data is sufficiently low. If the contribution to $D(g, y)$ is high, then smoothing should be 'on', or $e_{st} = 0$, to enforce a low value $(g_s - g_t)^2$.

Figs. 2.3 and 2.4 illustrate the paradigm of combining an edge model with a smoothing model. Figs. 2.3 displays MAP estimators for the models (2.4), (2.7), and (2.8). The differences are clearly visible in the profiles along straight lines. If we look at the two dimensional images thoughtfully and vary the distance to the picture we find that the visual impression of the pictures is not drastically different. This shows that the human imaging system works perfectly well and is robust against many kinds of distortions of usual images. It seems not to be trained to do the same with the one-dimensional plots. A similar effect is illustrated in Fig. 2.4 where 2D scenes are contrasted with 3D surface plots.

Remark 2.2.1. Perhaps it is helpful to think of the image g as an elastic membrane which is drawn towards data y by the term $D(g, y)$. In its neutral position the membrane is as flat as possible, ideally it is constant. Stretching increases the energy by the squared derivative; the discrete directional derivative $(g_s - g_t)/h$ between an s and its neighbour t at distance h is the elongation of the membrane. Because in our case $h = 1$, the membrane is elongated by the factor $|g_s - g_t|$ which then is squared. The important issue is that the membrane breaks if data enforce this and this entails energy costs α. This is the basic idea along which A. Blake and A. Zisserman developed their discrete version of the variational approach of D. Mumford and J. Shah, D. Mumford and J. Shah (1989), see their monograph A. Blake and A. Zisserman (1987). They do not work in the Bayesian framework and restrict themselves to the special case $D(g, y) = \sum_s (y_s - g_s)^2$, cf. Example 6.2.3. Independently S. Geman and D. Geman (1984) adopted similar models in the Bayesian formulation, which explicitly incorporated edge elements and allow for arbitrary noise, cf. the Examples 2.2.2 and 2.4.1.

Example 2.2.1. Let us compare the models (2.7) and (2.8). Let $S = \{1, 2, 3, 4\} \subset \mathbb{Z}$ with neighbour pairs $\{1, 2\}$, $\{2, 3\}$, and $\{3, 4\}$. Set all parameters β, λ^2, and α equal to 1. To avoid calculations, choose data $y_1 = -1/2 = y_3$, $y_2 = 1/2 = y_4$, and $g_i = 0$ for every i. Then $G^1(g, y) = 1 = L^1(g, y)$. This is a low value underlining the smoothing effect of both functions. Choose now data $y_1 = -2 = y_2$ and $y_3 = 2 = y_4$ with a jump between $s = 2$ and $s = 3$. For $g = y$ you get $G^1(g, y) = 16$ whereas $L^1(g, y) = 1$. Hence edge preservation is favourable for L^1 whereas it is penalized by G^1. Increasing the jump height increases this difference drastically.

In summary, the model (2.8) consistently combines smoothing and preservation of boundaries. Boundaries explicitly enter the model. There is a general

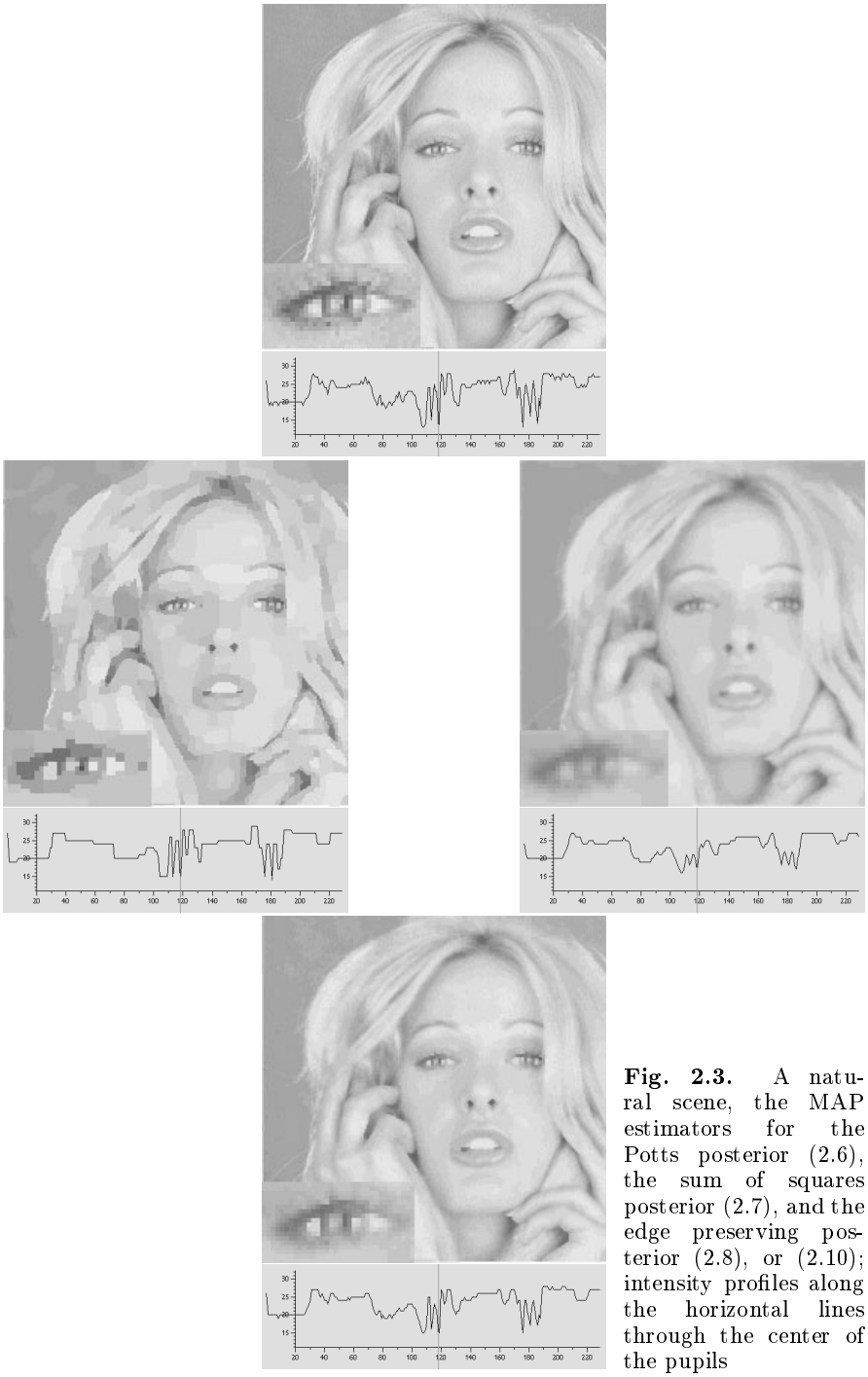


Fig. 2.3. A natural scene, the MAP estimators for the Potts posterior (2.6), the sum of squares posterior (2.7), and the edge preserving posterior (2.8), or (2.10); intensity profiles along the horizontal lines through the center of the pupils

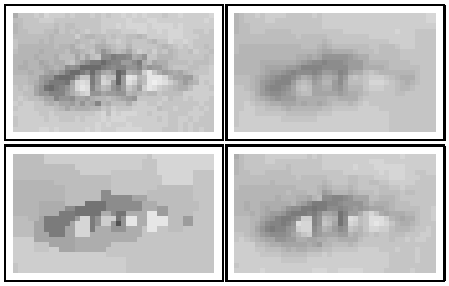
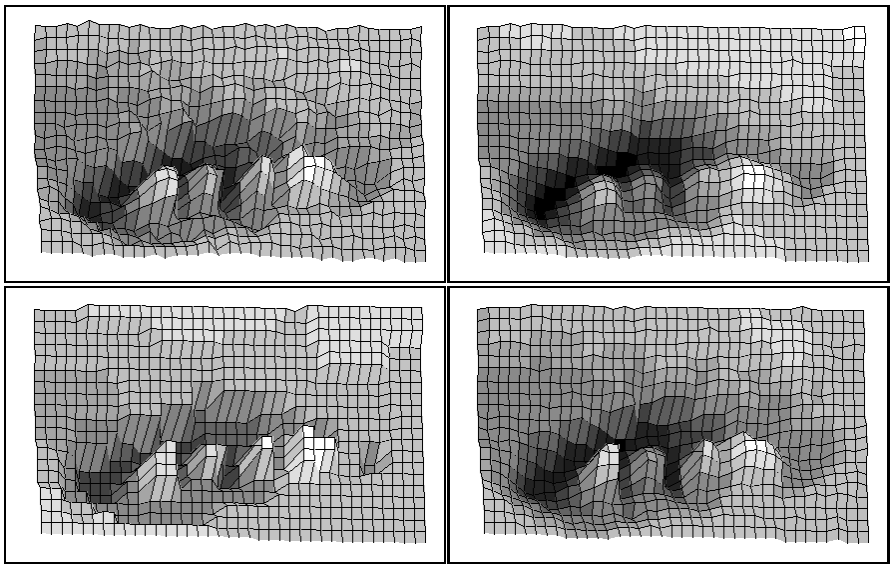


Fig. 2.4. Upper array: the girls eyes from Fig. 2.3; lower array: same as 3D-surface plots



principle behind this method. A homogeneous local operation is switched off where there is evidence that it does not make sense. Simultaneously, the set where the operation is switched off is organized according to regularity requirements; presently these are formulated in terms of the contour length. We shall meet this principle in various other models, for example in texture segmentation (Part V), or in motion analysis (Chapter 21).

The following simple but fascinating observation reveals that this is equivalent to a robustification of the prior distribution. What this means will be explained after some elementary calculations. Let (g^*, e^*) be a minimizer of $H(g, e)$. The data term does not depend on e and hence still has the form $D(g, y)$. We compute

$$\begin{aligned} H(g^*, e^*) &= D(g^*, y) + \sum_{s \sim t} \lambda^2 (g_s^* - g_t^*)^2 (1 - e_{st}^*) + \alpha e_{st}^* \\ &= \min_{g,e} \Big(D(g, y) + \sum_{s \sim t} \lambda^2 (g_s - g_t)^2 (1 - e_{st}) + \alpha e_{st} \Big) \end{aligned}$$

$$
\begin{aligned}
&= \min_g D(g,y) + \sum_{s\sim t} \Big(\min_{e_{st}=0,1} \lambda^2 (g_s - g_t)^2 (1-e_{st}) + \alpha e_{st} \Big) \\
&= \min_g D(g,y) + \sum_{s\sim t} \min\{\lambda^2 (g_s - g_t)^2, \alpha\} \\
&= \min_g D(g,y) + \sum_{s\sim t} \varphi(g_s - g_t), \qquad (2.9)
\end{aligned}
$$

where the function φ is the truncated square function (2.11). Hence an optimal intensity pattern g^* is a minimizer of the function

$$g \longmapsto \sum_{s\sim t} \varphi(g_s - g_t) + D(g,y), \qquad (2.10)$$

and vice versa. This function does not depend on e anymore. The component e^* of (g^*, e^*) can be uniquely reconstructed from g^* since:

$$e^*_{st} = 1 \Longleftrightarrow |g^*_s - g^*_t| > \delta = \sqrt{\alpha}/\lambda, \;\; \lambda > 0.$$

In this model intensity differences $|g_s - g_t| \geq \delta$ are treated like in the Potts model whereas differences $|g_s - g_t| < \delta$ are not recognized as jumps and hence smoothed according to (2.5). The message is that explicitly including edge elements and thereby modelling boundaries is equivalent to replacing the square function in (2.5) by the truncated square function φ in (2.11). Graph and contour lines of the function (2.10) for two pixels and Gaussian D is displayed in Fig. 2.6.

The form of φ and (2.9) tell a statistician that edge preservation is closely related to robust statistics. Let us briefly indicate what 'robust' means.

Remark 2.2.2 (Robust estimators). Robust statistics has several aspects. The oldest one was handling outliers in data. A crude but nevertheless frequently adopted practice is to remove what one believes to be outliers. Robust statistics deals with outliers without a need to identify them. We quote from F.R. Hampel et al. (1986):

> This 'classical approach' is founded on stringent stochastic models, and before long it was noticed that the real world does not behave as nicely as described by their assumptions. ... parametric models are used as vehicles of information, and procedures that do not depend critically on the assumptions inherent in these models are implemented.

In the simplest case one has i.i.d. random variables $Y_1, \ldots, Y_n$ with common finite expectation $\mathbb{E}(Y_i) = m$, and wants to estimate m. To this end one constructs an estimator or a statistic t which for each realization $y = y_1, \ldots, y_n$ returns an estimate $t(y_1, \ldots, y_n)$. A traditional procedure is to choose a minimizer of the sum of squares of residuals

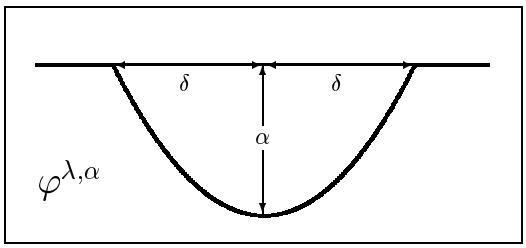


Fig. 2.5. The 'cup function'

$$\begin{aligned}\varphi(u) &= \varphi^{\lambda,\alpha}(u) \qquad (2.11)\\ &= \min\{\lambda^2 u^2, \alpha\}\end{aligned}$$

$$\vartheta \longmapsto L(y,\vartheta) = \sum_{i=1}^{n} \big(y_i - \vartheta\big)^2.$$

Setting the derivative with respect to ϑ to 0 gives the arithmetic mean $\bar{y} = (1/n)\sum_i y_i$ which is the *BLUE*, i.e. the best linear unbiased estimator. This means that among all *linear* estimators with expectation m it has the least variance. If the variables are Gaussian then it is even the best among *all* unbiased estimators, and simultaneously it is the maximum likelihood estimator, i.e. that value which maximizes the density (1.9). Could we wish for anything better? Unfortunately, the performance of the mean changes drastically if the model assumptions are violated. A single large deviation from m in $y_1, \ldots, y_n$ has a strong influence on L since it is weighted heavily by the squares, and hence on $\bar{y}$ itself, cf. Example 2.3.2.
Violation of the model assumption can appear in various ways. We may, for example, guess that the distribution is Gaussian, but in reality it is near the Laplace distribution, which has more heavy tails, cf. again Example 2.3.2. Another possibility is that the true distribution μ is close to the Gaussian $\mathcal{N}(m,\sigma^2)$ but there is *contamination* by some other distribution ν:

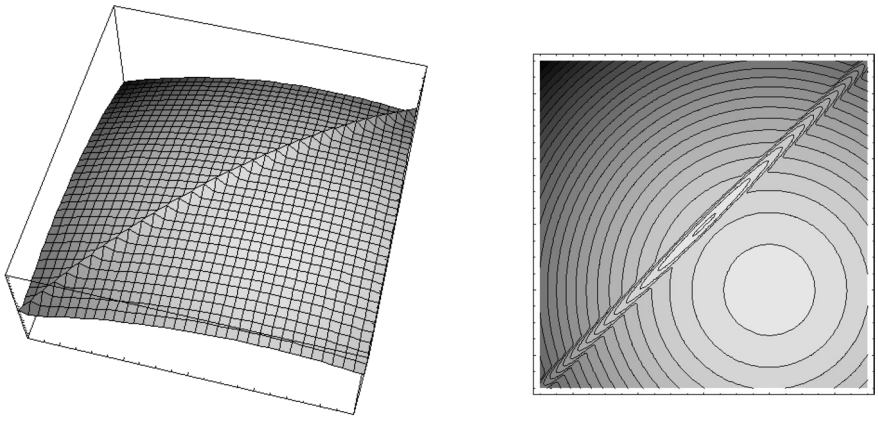


Fig. 2.6. Graph of (2.10) for two pixels 1 and 2: $(g_1, g_2) \mapsto \min\{100\cdot(g_1-g_2)^2, 10\} + (g_1-2)^2 + (g_2+2)^2$, seen from below and contour lines

$$\mu = (1-\varepsilon)\mathcal{N}(m,\sigma^2) + \varepsilon\nu,\ \ 0 \le \varepsilon \le 1.$$

Such a contaminated distribution may have heavy tails causing outliers. Insert for example $\nu = \mathcal{N}(m, 100\sigma^2)$ with a small ε. As a remedy one may replace the squares by functions penalizing large differences more moderately. This leads to *M-estimators*, which minimize functions

$$\vartheta \longmapsto \sum_{i=1}^{n} \varrho(y_i - \vartheta),$$

with $\varrho(u)$ increasing slower in $|u|$ than u^2, see PETER J. HUBER (1981). Least squares are the case $\varrho(u) = u^2$. HUBER suggests the least convex function which coincides with $\lambda^2 u^2$ in a ball; it has the form

$$\varrho(u) = \begin{cases} \lambda^2 u^2 & \text{if } |u| < \delta \\ 2\lambda\sqrt{\alpha}\,|u| - \alpha & \text{if } |u| \ge \delta \end{cases},\ \ \delta = \sqrt{\alpha}/\lambda. \tag{2.12}$$

The associated Gibbs distribution, for which the MAP estimator is the maximum likelihood estimator, has density $h(u) \propto \exp(-\varrho(u))$ and is called the *least favourable* distribution since in a neighbourhood of the normal law its maximum likelihood estimator has largest variance. In the M-estimator, ϱ reduces the influence of outliers but does not remove it completely.

To kill the influence of outliers completely, HAMPEL cuts off the branches of ϱ and introduces the φ-function

$$\varphi(u) = \begin{cases} \lambda^2 u^2 & \text{if } |u| < \delta \\ \alpha & \text{if } |u| \ge \delta \end{cases},\ \ \delta = \sqrt{\alpha}/\lambda,$$

see F.R. HAMPEL et al. (1986). We realize that this is precisely the function we derived by the calculations (2.9) from the edge model (2.8)! We can now argue the other way round: reading the calculations in reverse order shows that a radical robust approach to edge-preserving smoothing leads to the edge model (2.8). The interpretation from the robust point of view is quite natural: Consider a jump as displayed in Fig. 2.9. On the left half, variables have a law μ_a with mean a and on the right half the law is μ_b with mean b, $a \ne b$. As a neighbourhood or a window moves from left to right across the jump, there is more and more contamination of μ_a by μ_b until μ_a has been completely turned into μ_b.
Letting $\alpha \to 0$ and $2\lambda \to \infty$ such that $\lambda\alpha^{1/2} \to \gamma \in \mathbb{R}$ in HUBER'S proposal (2.12) gives a weighted modulus $\varrho(u) = \gamma|u|$. It leads to a robust L^1-theory of statistics; it is much harder than the usual L^2-theory for least squares, cf. P. BLOOMFIELD and W.L. STEIGER (1983). The corresponding L^1-prior $\mathit{\Pi}(x) \propto \exp(-\beta \sum |x_s - x_t|)$ is the most popular 'edge-preserving' prior. With the modulus we can play the same game as with the square and arrive at an edge type model with φ replaced by

$$\varrho(u) = \begin{cases} \lambda|u| & \text{if } |u| < \delta \\ \alpha & \text{if } |u| \ge \delta \end{cases},\ \ \delta = \alpha/\lambda.$$

For further discussion we refer to Example 2.3.3, G. WINKLER et al. (1999), and G. WINKLER and V. LIEBSCHER (2002).

Let us close this section with the classical example of a Bayesian edge model.

*Example 2.2.2 (*S. GEMAN and D. GEMAN *(1984)).* Bayesian imaging became popular in the statistics community after the seminal article S. GEMAN and D. GEMAN (1984), [130], see also D. GEMAN (1990). Their idea of piecewise smoothing by means of edge elements is nearly identical to that in A. BLAKE and A. ZISSERMAN (1987). The main difference is that the former authors adopt the Bayesian approach and thereby are free to incorporate statistical properties of noise and additional image features; the latter authors aim at the special GNC algorithm, cf. Section 6.2.3, which strictly limits their model to the form (2.8). The model in [130] also encapsulates various terms intended to capture regularity properties of boundaries. Our discussion above already included the regularity requirement of their shortness. In addition, in [130] selected undesired local edge configurations are penalized to control the shape of boundaries.
The context is the same as for (2.8). The state space is $\boldsymbol{X} = \boldsymbol{G} \times \boldsymbol{E}$ with the spaces $\boldsymbol{G}$ of intensity patterns and $\boldsymbol{E}$ of edge configurations. The joint prior distribution of g and e is given by:

$$\Pi(g,e) \propto \exp\big(-K(g,e)\big), \quad K(g,e) = K_S(g,e) + K_E(e).$$

The first term is responsible for boundary preserving smoothing and is very similar to that in (2.8). The second one controls the shape of boundaries. The smoothing term is given by

$$K_S(g,e) = \vartheta^2 \sum_{s\sim t} \psi\big(g_s - g_t\big)\big(1 - e_{st}\big). \tag{2.13}$$

The authors propose functions ψ similar to φ in (2.11), for example (2.15). Since $\lambda^2 u^2(1-v) + \lambda^2 \alpha v = \lambda^2(u^2-\alpha)(1-v) + \lambda^2\alpha$, and since addition of

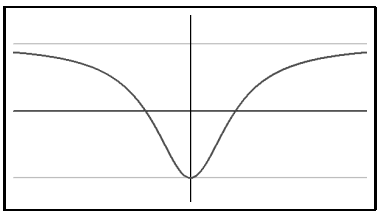


Fig. 2.7. Another cup function:

$$\psi(u) = 1 - \frac{2}{1 + (u/\delta)^2} \tag{2.15}$$

constants does not affect the Gibbs distribution, the model is of the same form as (2.8) but with ψ instead of $u^2 - \alpha$ and with $\alpha' = \lambda^2\alpha$ instead of α. In view of the discussion preceding (2.11), this amounts to a 'double robustification' since the edge terms correspond to a truncation of ψ and hence cancel the

effect '$\psi(u) \to$ constant' as $|u| \to \infty$. For small dynamic range - say up to 15 grey values - the authors recommend the Potts type function $\psi(0) = -1$ and $\psi(u) = 1$ otherwise. In summary, the discussion of this smoothing term is the same as above.
The function $K_E(e) = -\alpha W(e)$, $\alpha > 0$, serves as an additional organization term for the edges. W weights selected local edge configurations with a large factor if they are desired and with a small one if they are not. A sample of local configuration is displayed in Fig. 2.8. Edges should not be set inside smooth regions and therefore 'empty' local configurations (a) get large weights w_0. Smooth boundaries around smooth patches are welcome and configurations (b) are weighted by $w_1 < w_0$; sharp turns and T-junctions (c) and (d) get weights $w_3 \leq w_2 \leq w_1$ and blind endings and crossings (e) and (f) are penalized by weights $w_4 < w_3$. One may add an 'index of connectedness' and

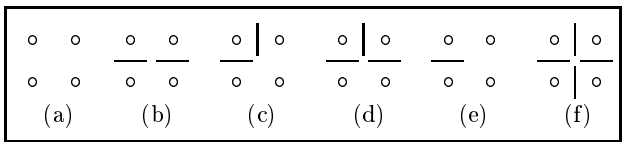


Fig. 2.8. Selected local configurations of active edges: nothing, straight line, T-junction, sharp turn, blind end, crossing

further organization terms; we postpone such supplements to Example 2.4.1. The prior energy function $K = K_S + K_E$ is specified now. Given a model for degradation and an observation y of degraded grey values the posterior can be computed (Example 1.2.5) and one can compute posterior estimates like MAP estimates. Note that posterior means do not give precise boundary values $e_{st} = 0, 1$. Nevertheless they may be used followed by rounding to 0 and 1.
This example illustrates a crucial aspect of contextual models. Smoothing, boundary finding, and organization of boundaries are simultaneous and cooperative processes.

2.3 Filters, Smoothers, and Bayes Estimators

Let us discuss the problem of smoothing or denoising and edge preservation from another point of view. Suppose we want to recover or *restore* an intensity pattern x from observed intensity data y. We invent a clever algorithm which produces the output $\hat{x}$ if fed with the input y. For such a map $y \mapsto \mathcal{F}(y) = \hat{x}$ the designation *filter* is borrowed from engineering. Statisticians call it an estimator, at least if randomness is involved. In this sense all Bayes estimators introduced in Section 1.4 may be viewed as instances of filters.

Most conventional and also many recently developed filters act on signals taking values in Euclidean spaces $\mathbb{R}^d$ and not on finite discrete sets. In order

to compare Bayesian estimators with such methods, we must also let them take continuous values. Let us hence consider signals or images as elements of a space $\mathbb{X} = \{(y_s)_{s\in S} : y_s \in \mathbb{R}\}$. We use the symbol $\mathbb{X}$ to distinguish continuous signals from discrete ones, where we usually wrote $\boldsymbol{X}$.

Let us start now naively from the very beginning. In the signal analysis community, *linear filters* are very popular. A map $\mathcal{F}$ from $\mathbb{X}$ to $\mathbb{X}$ is called *linear* if it fulfills $\mathcal{F}(\alpha y+\alpha' y') = \alpha\mathcal{F}(y)+\alpha'\mathcal{F}(y')$ for all $y, y' \in \mathbb{X}$, and $\alpha, \alpha' \in \mathbb{R}$. One reason for the popularity of linear filters is Fourier analysis, which simultaneously is a powerful tool for their analysis, and a useful instrument for their practical implementation. The first simple example is concerned with the most frequently used filters.

Example 2.3.1 (Moving averages). *Moving averages* convolve the observed image with 'noise cleaning masks'. In the two-dimensional case, the latter are matrices $M = \big(M(k,l)\big)_{k,l=-q}^{q}$ such that the weights $M(k,l)$ are nonnegative and add up to 1; they were introduced in Example 1.2.5. On signals $y \in \mathbb{R}^{\mathbb{Z}^d}$ their action is

$$(My)_{(i,j)} = \sum_{k,l=-q}^{q} M(k,l)y_{(i-k,j-l)}, \tag{2.16}$$

where $(i,j) = s$ denotes a generic lattice point. If S is a finite lattice then the definition is modified near the boundary, or better, the upper and lower rim, as well as the left and right rim, are identified, and one works on a torus. Typical instances are the uniform and the binomial masks in (1.11). A large variety of such masks (and combinations) can be found in the toolbox of image processing. A classical reference for such filters is W.R. Pratt (1991), see also B. Jähne (2002), [211], or B. Jähne (1993) in German. The uniform filter usually over-smoothes; even worse, inspection of its Fourier transform shows that it does not remove roughness of certain 'wave lengths' (apply it to vertical or horizontal stripes of different width, cf. [211]). The Binomial filter performs much better but there is still oversmoothing. The filters frequently are iterated several times. Note that in Example 1.2.5 we used the same construction to model blur. In fact, such filters smooth or blur the signal in order to reduce the noise contribution. This works on flat parts, but jumps are more or less destroyed. This is illustrated in Fig. 2.9.

A typical example of a nonlinear filter is the moving median.

Example 2.3.2 (Moving median). For $s = (i,j)$ let $B_s = \{t = (i,j) \pm (l,k) : |l|, |k| \le q\}$ be windows, and observations y_t, $t \in B_s$, with values in $\mathbb{R}$, or an ordered space be given. Write the observations in the window in increasing order $y_{(1)} \le \cdots \le y_{(n)}$, where $n = (2q+1)^2$. The *moving median filter* is defined by $(\mathcal{M}y)_s = y_{((n+1)/2)}$, i.e. it takes as value in s the $y_{(k)}$ in the middle. If, for example $y_{-2} = 2, y_{-1} = 2, y_0 = 3, y_1 = -1, y_2 = 0$ then $(\mathcal{M}y)_0 = 2$ and the mean is 1.2. This median filter is much more robust against outliers than moving averages. Whereas the mean of $2, 1000, 3, -1, 0$

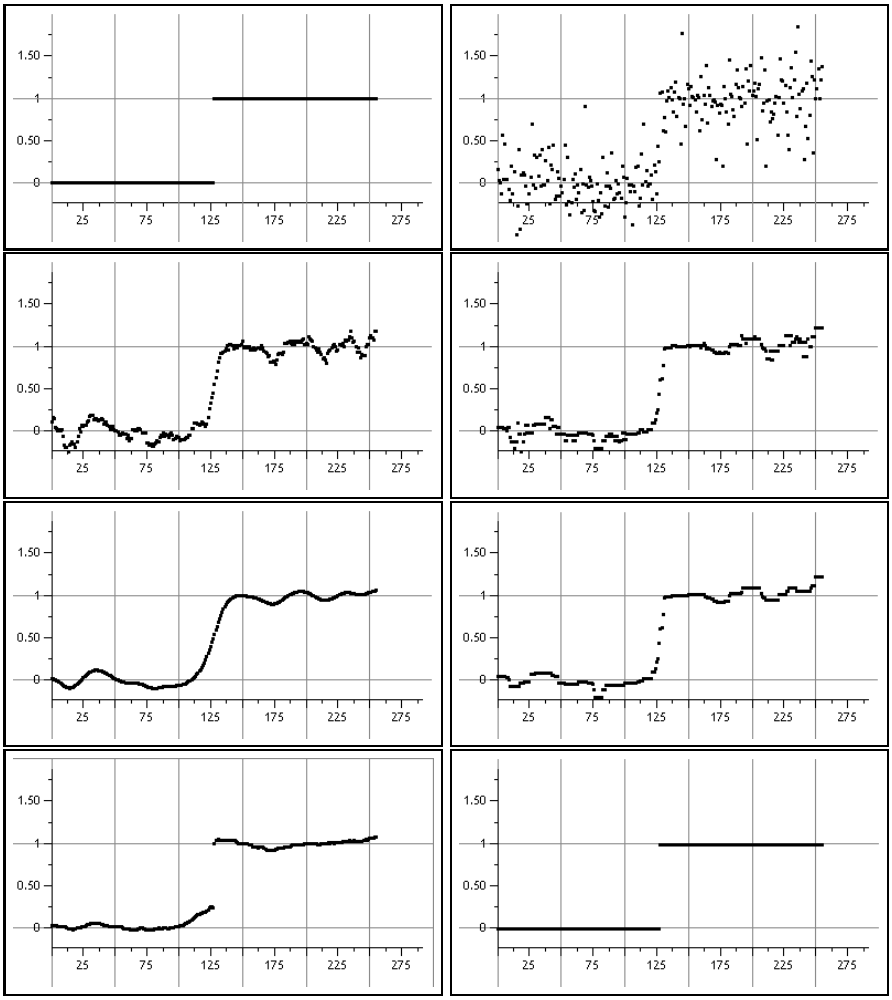


Fig. 2.9. Upper row: a jump of height 1, degraded by additive white Laplacian noise of standard deviation $\sigma = 0.2$; filtered with masks of length $2q + 1$, $q = 5$, once in the second, and five times in the third row. A uniform moving average in the left and the median filter in the right column. In the 4^{th} row on the left a MAP for the truncated square model with $\delta = 0.2$ and $\alpha = 10$, on the right for the Potts model with $\gamma = 0.1$. The error of the Potts MAP is about 1%

is $1004/5 \sim 201$, the median again is 2. The action of the moving median is illustrated in the right column of Fig. 2.9. The median here serves just as a simple example. Much more important in practice are *morphological filters* which are also based on the order of values. In particular, they are *idempotent*, i.e. fulfill $\mathcal{F} \circ \mathcal{F}(y) = \mathcal{F}(y)$ for every $y \in \mathbb{X}$. The median is related to them

but not morphological since it clearly is not idempotent. Standard texts on mathematical morphology are the monographs by J. SERRA (1982, 1988).

Let us compare a typical linear with a typical nonlinear filter. The simplest criterion for noise reduction is reduction of variance on a noisy flat region.

Example 2.3.3 (Noise reduction and robustness). The *noisy flat* simply is a collection Y_s, $s \in S$, of independent and identically distributed random variables; in this example we let them be centred with standard deviation 0.2. We will compare the moving median with the moving average with uniform weights, defined on the left hand side of (1.11).
Consider first variables $Y_1, \ldots, Y_n$ uniformly distributed on $[-\sqrt{3}/5, \sqrt{3}/5]$ such that $\sigma = 0.2$. Then the variance of the median is $\mathbb{V}(\mathcal{M}) = (3\sigma^2)/(n+2)$, whereas $\mathbb{V}(\bar{Y}) = \sigma^2/n$. For a 5×5-window with $n = 25$ we have $\mathbb{V}(\bar{Y}) = \sigma^2/25$ and $\mathbb{V}(\mathcal{M}) = \sigma^2/9$. The average performs much better than the median. For Gaussian noise we get

$$\mathbb{V}(\bar{Y})/\mathbb{V}(\mathcal{M}) = (2/\pi)(1 + (\pi - 2)/(2n)).$$

For the same mask, the variance of the median is about 57% larger than that of the average. Still the average is better than the median, but the median catches up. For Laplacian noise, which has a considerably heavier tail, we get

$$\mathbb{V}(\mathcal{M}) \approx \frac{\sigma^2}{2(n - 1/2)}$$

which is about half the variance of an average, and now the median is already superior to the average.

Since the variables in a moving window are i.i.d. these considerations apply to the moving average and the mowing median.

This reveals the message: the median becomes more and more superior to the average the heavier the tail of the noise distribution is. This means that 'the median is more robust than the average'. The shape and a portion of the tails are displayed in Figs. 2.10 and 2.11. For the proofs we refer to B.I. JUSTUSSON (1981) (German readers may also consult [341]).

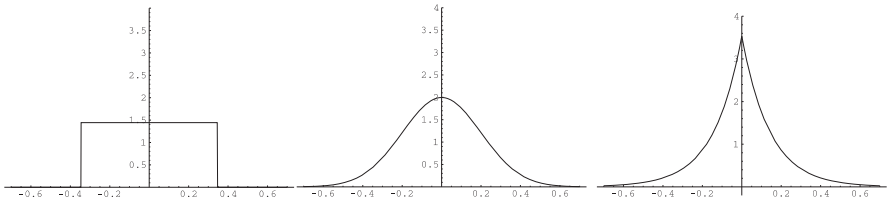


Fig. 2.10. Densities of the uniform-, Gaussian- and Laplace distribution with standard deviation 0.2. For this standard deviation, the uniform distribution is concentrated on the interval $[-\sqrt{3}/5, \sqrt{3}/5]$

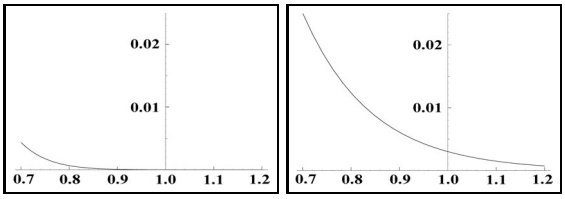


Fig. 2.11. Parts of the tails of the Gaussian and the Laplace distribution with standard deviation 0.2

To allow for edge preservation we extend the classical concept of a linear filter. A natural generalization is to have filter weights depending on the input: A map $\mathcal{F} : \mathbb{X} \to \mathbb{X}$ will be called a *convex filter* if for each $y \in \mathbb{X}$ and each site $s \in S$ there are weights $W_{st}(y)$ with

$$(\mathcal{F}y)_s = \sum_{t \in S} W_{st}(y) y_t, \;\; W_{st}(y) \geq 0, \;\; \sum_t W_{st}(y) = 1.$$

This generalizes the representation $\mathcal{F}y = Ay$ of linear filters by (stochastic) matrices A. Convexity means that the filter does not extend the range of the input:

Lemma 2.3.1. *A filter $\mathcal{F}$ is convex if and only if*

$$\min\{y_t : t \in S\} \leq (\mathcal{F}y)_s \leq \max\{y_t : t \in S\}.$$

Clearly, linear filters are convex because of their very definition, and median filters are convex by Lemma 2.3.1, since they just rearrange values.

Unfortunately, it turns out that for the models we discussed previously, prior distributions do not exist in the continuous setting. The reason is that most priors depended on differences $x_s - x_t$, these are not affected by the addition of constants, i.e $K((x_s + c)_s) = K((x_s)_s)$, if $c \in \mathbb{R}$, and hence the associated prior density $\Pi(dx) \propto \exp(-K(x))\, dx$ on $\mathbb{R}^S$ would be *translation invariant*. One can easily show that there is no probability density on $\mathbb{R}^S$ proportional to $\exp(-K(x))$. Nevertheless, under suitable integrability conditions measures corresponding to posterior distributions can be defined for many models by $\Pi(dx|y) \propto \exp(-K(x) - D(x,y))\, dx$. If this works then one still may speak about (pseudo) MAP, MMS, etc. estimators. For the Potts model these considerations are not meaningful since its prior energy function vanishes except on a set of Lebesgue measure zero. The pseudo MAP exists also for this model since no substitute for the posterior is necessary and one simply minimizes the function $x \mapsto K(x) + D(x,y)$. Notwithstanding the formal incompatibility, one can learn a lot about Bayesian models from their counterparts on continuous spaces.

For us it was a surprise that most MAP estimators are convex filters in this sense. Typically, there are functions ψ, ϱ and v such that

$$K(x) = \sum_{s,t} \psi(x_s - x_t) v(s-t), \;\; D(x,y) = \sum_{s \in S} \varrho(x_s - y_s). \tag{2.17}$$

In G. Winkler and V. Liebscher (2002) we show:

Proposition 2.3.1. *Let D and K be given by (2.17). Assume $v \geq 0$, that $\psi(u)$ and $\varrho(u)$ are symmetric around zero and increasing in $|u|$, and suppose further that ϱ is strictly increasing on $[0, \infty)$. Then each MAP-estimate is a convex filter.*

This holds in most practical cases. In this text v usually is the indicator function of a neighbourhood of $0 \in \mathbb{Z}^S$.

Remark 2.3.1. A lot of other recent nonlinear methods fit into this conception too. Examples closely related to the present discussion (G. WINKLER et al. (1999), G. WINKLER and V. LIEBSCHER (2002)) are the nonlinear Gaussian filter in F. GODTLIEBSEN et al. (1997), the nonlinear Gaussian filter chain in V. AURICH and J. WEULE (1995), [10], the local M-smoother in C.K. CHU et al. (1998), [63], and the adaptive weights smoother in J. POLZEHL and V.G. SPOKOINY (2000).

Edge preserving filters have to preserve significant intensity contrast. They should even decide whether a place is an intensity jump and perhaps mark the location - say by an active microedge. This raises the questions: 'What is a jump?' and 'Can we decide whether a method finds jumps, or not?'. The Potts model and the robust models (2.8) and (2.10) both include a precise criterion for what they declare to be a jump and where they locate it: for the latter a jump is between two neighbours s and t with contrast $|x_s - x_t| \geq \delta$ and for the former with contrast $|x_s - x_t| \neq 0$. In the global smoother (2.7) and in linear filters such criteria are not incorporated and they cannot decide upon jumps. A decision can be taken only after a subsequent nonlinear operation like thresholding.

The following sketches may shed some light on this problem. Since evident jumps are supposed to be large, we suggest the crude criterion

$$\lim_{r \to \infty} r^{-1} \mathcal{F}(ry) = y, \quad y \in \mathbb{X} \tag{2.18}$$

where $(ry)_s = ry_s$. No matter how small contrast may be we can force it to be a locally significant jump by multiplication by a large scale parameter r. Jumps in this sense vanish if r tends to zero. This concept captures jumps of fixed size. It does not include a notion of jumps which are large *relative* to the signal size.

Example 2.3.4 (Scale invariant filters). Consider now a linear filter $\mathcal{F}y = Ay$ with a stochastic matrix A. Then the identity $Ay = r^{-1}Ary = y$ shows that $\mathcal{F}$ fulfills (2.18) if and only if A is the identity. This may be rephrased as: 'Linear filters do not preserve edges'. The moving median is sometimes believed to preserve edges. In the sense of (2.18) it does not, since $r^{-1}\mathcal{M}(ry) = \mathcal{M}(y)$ and $r^{-1}\mathcal{M}(ry) = y$ for every y would again imply that $\mathcal{M}$ is the identity. This is the case if and only the mask contains precisely one site. This observation may at first glance look surprising. But we must keep in mind that the median only sees the ordering and does not feel scale. If we look at slowly varying parts

of the output through a magnifying glass it will look like the output in rough parts. Observe that for convex filters $\mathcal{F}$ constant signals y are preserved; in particular, $r^{-1}\mathcal{F}(ry) = y$ if y is constant.
More generally, call a filter *scale invariant* if $r^{-1}\mathcal{F}(ry) = \mathcal{F}(y)$. We saw that moving averages and medians are scale invariant. Generalizing, we may say that 'scale invariant filters do not preserve edges'.

The maximum posterior estimators for the Potts model and the robustified square (2.8) or (2.10) fulfill (2.18):

Proposition 2.3.2 (V. Liebscher (2002)). *Let*

$$H(x) = \sum_{s,t} \psi(x_s - x_t)v(s-t) + \sum_s \varrho(x_s - y_s).$$

Assume $v \geq 0$, that $\psi(u)$ and $\varrho(u)$ are symmetric around zero and increasing in $|u|$, and suppose further that ϱ is strictly increasing on $[0,\infty)$. Moreover assume that

$$\limsup_{|r|\to\infty} \frac{\psi(ru)}{\varrho(r)} = 0, \text{ for every } u > 0.$$

Let $\mathcal{F}(y)$ denote a corresponding MAP estimator. Then $\mathcal{F}$ fulfills (2.18). This holds in particular for the Potts prior $\psi(u) = 1 - \delta(0,u)$, or if ψ has a cup shape like in (2.11).

If ϱ is quadratic, then the condition that ψ increases not slower than ϱ is fulfilled by all ψ increasing not faster than linear. Examples for such functions are the robust cup-shaped functions φ and Huber's function.

Similar results hold for some other recent smoothers as well:

Remark 2.3.2. The Gaussian nonlinear sigma-filter from V. Aurich and J. Weule (1995) and the local M-smoother from C.K. Chu et al. (1998) fulfill (2.18) as well.

In the above discussion we had continuous intensities but discrete space. There is also a theory for continuous space which basically deals with Mumford-Shah energy functionals. These are formally similar to the discrete edge model (2.8). The discrete arrays (x_s) and (y_s) are replaced by functions $x(u)$ and $y(u)$ on some domain $D \subset \mathbb{R}^d$, with certain regularity properties. There are subsets K of D enclosing points of irregularity, for example points where the functions $x(u)$ are allowed to be discontinuous across K; these are interpreted as locations of boundaries. For $d = 2$, the edge set K is assumed to be some curve with length $L(K)$. The original functional has the form

$$\mathcal{E}(x,K) = \lambda^2 \int_{D\setminus K} \left\|\nabla x(u)\right\|_2^2 du + \alpha \cdot L(K) + \int_D \left|x(u) - y(u)\right|^2 du$$

This is the continuous counterpart from D. MUMFORD and J. SHAH (1989) to the discrete model (2.8) (nearly everywhere one finds the citation D. MUMFORD and J. SHAH (1985), but this paper seems never to have appeared. We found contradictory comments). E. DE GIORGI (1991) adopts the modified version

$$\mathcal{E}(x) = \lambda^2 \int_D \left\|\nabla x(u)\right\|_2^2 du + \alpha \mathcal{H}^{d-1}(S_x) + \int_D \left|x(u) - y(u)\right|^2 du.$$

The main difference is that the $d-1$-dimensional set S_x of discontinuities is weighted by the $d-1$-dimensional Hausdorff measure $\mathcal{H}^{d-1}$. A. BLAKE and A. ZISSERMAN (1987) discuss their Graduated Nonconvexity Algorithm for the actual computation of minimizers for discrete space. The state of the art in 1995 is reported in J.-M. MOREL and S. SOLIMINI (1995), a more recent account is L. AMBROSIO et al. (2000).

2.4 Boundary Extraction

Edge detection or boundary finding is an important field of image analysis. Edges correspond to sudden changes of an image attribute such as luminescence or texture and indicate discontinuities in the actual scene. There is an enormous variety of filtering techniques for edge detection. Most are based on discrete derivatives, frequently combined with smoothing at small scale to reduce the noise contribution. There are also many ways to do some cosmetics on the extracted raw boundaries, for example erasing loose ends or filling small gaps. More refined methods like fitting step shaped templates locally to the data have been developed, cf. the monographs H. NIEMANN (1990), A. BLAKE and A. ZISSERMAN (1987).

The following Example 2.4.1 is of historical interest, since it is one of the first which goes beyond filtering and models the shape of boundaries in the Bayesian context. It is reported in D. GEMAN (1987) and D. GEMAN et al. (1987). It is closely related to Example 2.2.2; the new idea is to model boundaries in their own right and not merely as strings of active microedges.

Example 2.4.1 (Boundary extraction). We continue with notation from Example 2.2.2. Recall that microedges were virtual edges $s \sim t$ between neighbour pixels s and t. In Section 1.1 and in Example 2.2.2, we identified boundaries with the set of active microedges, i.e. neighbour pairs s, t with $e_{st} = 1$. Albeit boundaries are defined by means of edges in the present example, they are now image features in their own right.
The locations of boundary elements will be between those of edge elements; instead of a formal definition we indicate this in Fig. 2.12. Let the set of these boundary locations '$*$' be denoted by B. Then, we can define active and inactive boundary elements, and respectively let $b_u = \pm 1$, $u \in B$.

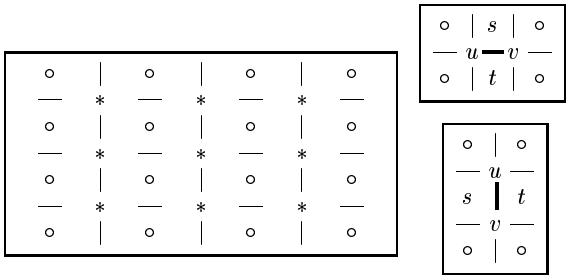


Fig. 2.12. Positions of pixels, microedges and boundary elements, left:
∘ pixel site
| microedge site
∗ boundary location.
Right: Adjacent boundary locations $u \sim v$

Hence we have a state space $\boldsymbol{X} = \boldsymbol{G} \times \{0,1\}^B$ containing intensity patterns g and boundary configurations b. Similar to Example 2.2.2 the joint prior distribution between intensities and boundaries is given by

$$\Pi(g,b) \propto \exp\big(-K(g,b)\big), \;\; K(g,b) = K_S(g,b) + K_B(b). \tag{2.19}$$

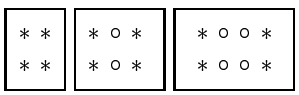
The first term K_S is responsible for seeding boundaries, and the second term K_B provides boundary organization in accordance to our geometrical and topological expectations. Seeding is based on contrast and continuation:

$$K_S(g,b) = \vartheta_1 \sum_{u \sim v} \psi\big(\delta_{u,v}\big)\big(1 - b_u b_v\big) + \vartheta_2 \sum_{u \in B} \big(b_u - \zeta_u(g)\big)^2, \;\; \vartheta_1,\, \vartheta_2 > 0.$$

In the left term, summation extends over pairs of adjacent boundary locations $u \sim v$. In between there are microedges | , or — , separating pixels $s(u,v) \sim t(u,v)$. $\delta_{uv}(g)$ is the contrast $|g_{s(u,v)} - g_{t(u,v)}|$ across this microedge. Again, ψ is an increasing function of contrast; in [127] the authors use $\psi(\delta) = \delta^4\big(c + \delta^4\big)^{-1}$, but the essential point is that ψ has a shape similar to φ from (2.11). The right term depends on an index $\zeta(g)$ of connectedness: given thresholds $c_1 < c_2$, a microedge is called active if either (i) the contrast across the microedge exceeds c_2 or (ii) the contrast exceeds c_1 and the contrast across one of the neighbour microedges exceeds c_1. The index $\zeta_u(g)$ equals 1 if u is inside a string of say four active microedges and 0 otherwise.
The second term in (2.19) organizes the boundary configuration b:

$$K_B(b) = \vartheta_3 \sum_{C \in \mathcal{C}} \prod_{u \in C} b_u - \vartheta_4 W(b), \;\; \vartheta_3 > 0,\; \vartheta_4 > 0.$$

The first term penalizes double boundaries counting the local boundary configurations depicted in the little plots (and their rotations by 90 degrees).



('∗' means that there is a boundary element and '∘' that there is none). The members C of $\mathcal{C}$ are the corresponding sets of boundary sites. Like in Example 2.2.2, the second term penalizes local configurations. Be aware that this is an example from 1987. Nevertheless it is a model in which

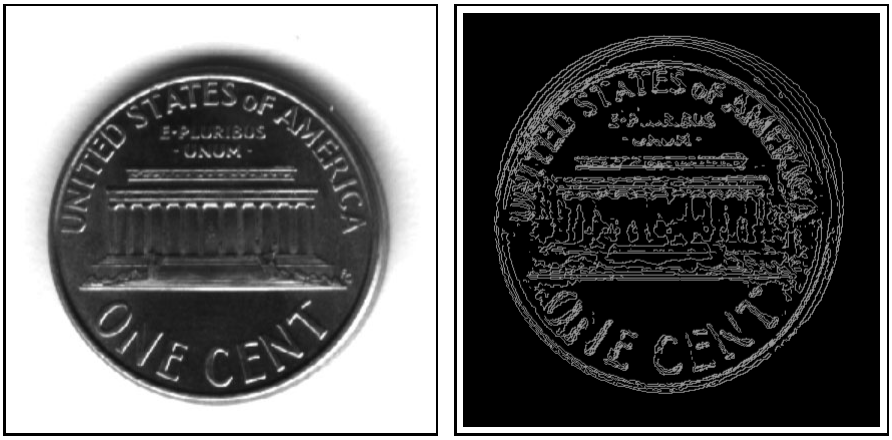


Fig. 2.13. Left: Original image, right: Boundaries estimated by rounding an MMS estimate from an edge model from Example 2.4.1

the processes of seeding and organization are entirely cooperative. Low contrast segments may survive if sufficiently well organized and, conversely, unstructured boundary segments are removed by the organization terms.
The right picture in Fig. 2.13 shows a thresholded MMS estimate for such a boundary model.

2.5 Dependence on Hyperparameters

Another important aspect is model choice. We mentioned already that configurations x correspond to the quantities called parameters in Bayesian statistics. In practically all models there are additional parameters of another type, like λ and α in (2.10) or β in (2.4) or (2.7). They are called *hyperparameters*. In our previous considerations they were part of a fixed model and assumed to be known. It is obvious that estimators crucially depend on these hyperparameters; Fig. 2.16 illustrates this impressively. And precisely there is the rub! Although we may have good reasons and a precise idea of the general form of the prior, we nevertheless may have no idea about hyperparameters appropriate for a special data set. In many articles, they are chosen by trial and error, and in others ad hoc methods are invented, cf. [130]. Frequently another prior is put on the hyperparameters. Hyperparameters are one of the greatest obstacle to be removed in order to turn a Bayesian method into a practicable algorithm. Let us illustrate dependence on hyperparameters by way of an example which is simple, of practical relevance, and gives us the opportunity to comment further on modelling.

Example 2.5.1 (Hyperparameters in the Potts model). The data displayed as dots in Figure 2.16 are measurements from fMRI (functional magnetic

resonance imaging or tomography) of the human brain. The aim is to identify regions of the brain responding to an outer stimulus by increased activation. Increased activity causes increased metabolic rate which in turn is followed by increased delivery of blood to the activated region. The measurements are based on the BOLD imaging technique. It does not measure tissue perfusion or flow directly, however, but depends on the presence of blood deoxygenation and that deoxigenated haemoglobin is a *blood oxygen level dependent* effect (BOLD) that can be observed by noninvasive magnetic resonance imaging at high magnetic fields.
In the experiment a person is exposed to a visual stimulus; in the present case a chequerboard pattern was periodically switched on and off. This 'on and off' is a signal of boxcar type like in Fig. 2.14. One expects that in

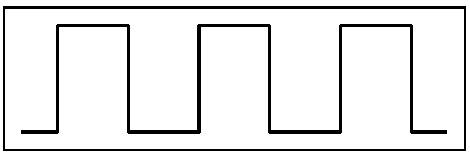


Fig. 2.14. A boxcar signal

certain brain regions, for example in the visual cortex, neurons respond to this stimulus with a signal of similar shape, which can be recorded by means of fMRI. Data in Fig. 2.16 show a time series of 70 measurements in a *voxel* of the visual cortex (a voxel is a three-dimensional pixel) of about $3 \times 3 \times 5$ mm^3. The question is wether the time series shares crucial features with the boxcar stimulus and can be labelled as 'activated'.
In the present context, we use these data to illustrate dependence of MAP estimators on hyperparameters. We adopt a one-dimensional Potts model with $S = \{1, \ldots, n\}$ and neighbours $i \sim i+1$. We write it in the form

$$L^\gamma(x, y) = \gamma \big| \{ i : x_i \neq x_{i+1} \} \big| + \sum_{i=1}^{n} (x_i - y_i)^2, \;\; x_i \in \mathbb{R}, \;\; \gamma \geq 0. \qquad (2.20)$$

We want to illustrate how strongly MAP estimates, i.e. minimizers x^*_γ of the function $x \mapsto L^\gamma(x, y)$, depend on γ.
Note that the signal x in this example takes values in $\mathbb{R}^n$ and not in a finite discrete set. This has two reasons: the continuous case is analytically easier to handle, and MAP estimators can be computed *exactly* by means of dynamic programming. Exact MAPs are necessary to ensure that the observed dependence of estimates on hyperparameters is not obscured by inaccurate computations.
It is clear that the MAP estimator for (2.20) returns the original data if $\gamma = 0$ and a constant x^* for $\gamma = \infty$. It is also plausible that the number of jumps in the MAP estimate should increase as γ decreases. One can show more.

Proposition 2.5.1. *There is a set $N \subset \mathbb{R}^n$ of Lebesgue measure zero such that for each $y \notin N$ there is a sequence $\gamma_1 > \cdots > \gamma_m > 0$ such that the following holds:*
(a) For each γ in the intervals (γ_1, ∞), (γ_{k+1}, γ_k) and $(0, \gamma_m)$ there is a unique minimizer x_k^* of (2.20).
(b) For all $\gamma > \gamma_1$ the MAP estimate x_0^* is a constant time series, and for $0 \le \gamma < \gamma_m$ one has $x_m^* = y$.
(c) For $\gamma = \gamma_k$ the only two minimizers of (2.20) are x_{k-1}^* and x_k^*.
(d) The number of jumps of the x_k^* increases strictly in k.

This is shown in A. KEMPE (2003). Part (a) of the proposition means that for every density on $\mathbb{R}^n$ with respect to Lebesgue measure almost all y have a unique MAP estimate. Fig. 2.15 illustrates such γ-intervals. Let us check

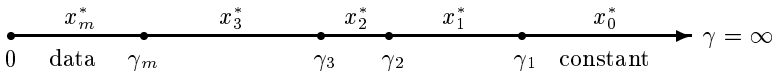


Fig. 2.15. On the intervals (γ_{k+1}, γ_k) the MAP estimate does not change

now how Proposition 2.5.1 works on the brain data. Fig. 2.16 shows the MAP estimates based on the (dotted) data y for the first six subsequent γ-intervals

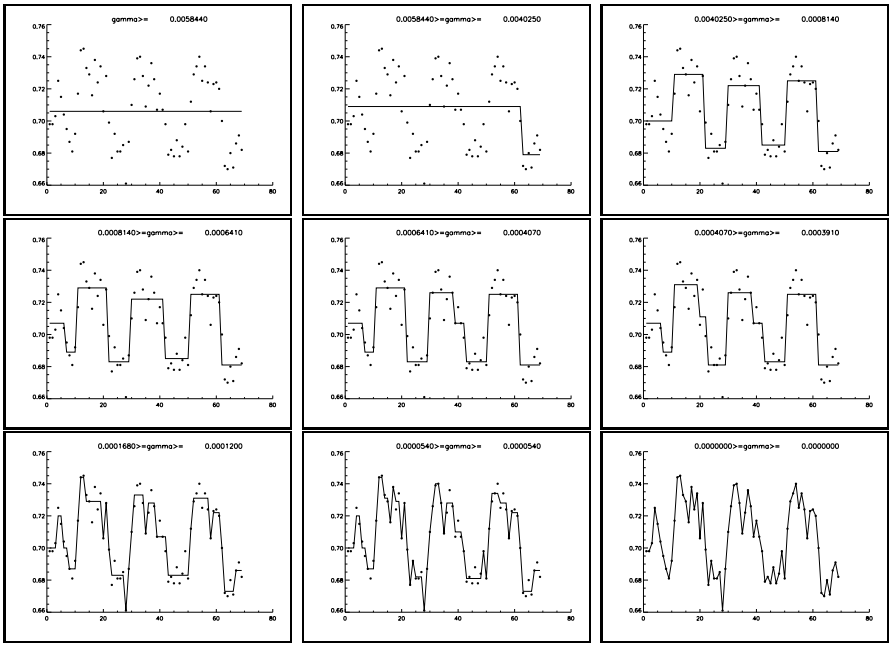


Fig. 2.16. Brain data, and MAP estimates on γ-intervals

beginning from the right. It also displays the MAPs for the 18^{th}, 25^{th}, and 51^{st} of the 51 γ-intervals, including $[\gamma_1, \infty)$. The MAP estimates were computed *exactly* by means of dynamic programming (which works without problems in one dimension and on trees, cf. [344; 229]).
It is interesting to have a closer look at the lengths of these γ-intervals. We have $\gamma_1 \approx 58 \cdot 10^{-4}$. The three modes of the stimulus appear first in $[\gamma_3, \gamma_2)$ with $\gamma_2 \approx 40 \cdot 10^{-4}$ and the MAP estimate basically stays unchanged until $3.9 \cdot 10^{-4}$ (in statistics a *mode* is a local maximum). Hence in the present example, this 'stable' γ-region fills more than 94% of the interval $[0, \gamma_1)$. This is a strong indication that we should search for the correct hyperparameter in this region. This is work in progress, cf. A. KEMPE (2003).

Inspired by Example 2.14 we bring a further aspect of probabilistic modelling up for discussion. In this - and in many, if not most, real applications, at least in the life sciences - the (random) transformation of the 'true' object x to data y basically is unknown. It is impossible to formalize precisely all the way from the visual stimulus through receptors, nerve paths and all the biochemical reactions there, neurons, request for oxygen, magnetic fields, excitation of spins, their relaxation times, and their measurement by a complicated technical machinery. There is nothing else we can do than to simplify matters drastically. To stick to our example, we may ask what a response to the stimulus should be. Differently phrased we should formulate *minimal qualitative criteria* as a basis for statistical decisions. P.L. DAVIES (1995) calls this *parsimonious statistics*. For the brain data we might for example decide whether neurons in a voxel respond or not comparing the number of significant modes in the stimulus and the response, cf. Figs. 2.14 and 2.16.