On the Equivalence of Local-Mode Finding, Robust Estimation and Mean-Shift Analysis as used in Early Vision Tasks

Rein van den Boomgaard Joost van de Weijer

Intelligent Sensory Information Systems Group
University of Amsterdam, The Netherlands
E-mail: {rein,joostw}@science.uva.nl

Abstract

In this paper we show the equivalence of three techniques used in image processing: local-mode finding, robust-estimation and mean-shift analysis. The computational common element in all these image operators is the spatial-tonal normalized convolution, *an image operator that generalizes the bilateral filter.*

1. Introduction

The luminance values in a local neighborhood are often from two distributions: the foreground distribution and the background distribution. Calculating the average of all luminance values is therefore bound to smooth the boundaries of the depicted objects.

The histogram of the luminances in the neighborhood of a point will show two peaks. Finding the mode (local maximum) in the histogram that is most likely to represent the distribution that the point belongs to, leads to local mode filtering, a technique that is shown to result in visually impressive results (see van de Weijer et. al. [9]).

A way to circumvent mixing the values from foreground and background distributions is to consider some of the values found in the local neighborhood as statistical outliers. Robust estimation of local image structure has been used in the past (see Besl et. al. [1]). In recent years robust estimators in connection with non-linear diffusion techniques have been used successfully (see Black et. al. [2]).

The connection between mean-shift analysis and local-mode estimation is known for some time (see Cheng [3]). Mean-shift analysis is used with success in several early vision tasks (see [4]). The connection between mean-shift analysis and robust estimators has not been reported in literature to the best of our knowledge.

In this paper we show the equivalence of these three techniques (local-mode finding, robust-estimation and mean-shift analysis) as they are used in early vision. The computational common element in all these image operators is the *spatial-tonal normalized convolution*, an image operator that generalizes the bilateral filter introduced by Tomasi and Manduchi [8].

The connection with local-mode estimation firmly sets our work in the context of locally orderless images as they are introduced by Grifin [5] and Koenderink and Van Doorn [7]. The connection with robust estimation of local image structure allows us to look at higher order image structure as well. Finally the connection with mean-shift analysis brings in a wealth of results concerning the well-posedness, stability and accuracy of the iterative numerical schemes that are used.

2. Spatial-Tonal Normalized Convolution

Tomasi et al.[8] introduced the bilateral filter as an intuitive appealing generalization of the (Gaussian) convolution. In this section we follow their line of thought to introduce the *spatial-tonal normalized convolution* that is an generalization of the bilateral filter.

Low pass filtering (smoothing) of an image f results in the normalized convolution:

$$\frac{\int_{\mathbb{R}^d} f(\mathbf{y})\, v(\mathbf{x}-\mathbf{y})\, d\mathbf{y}}{\int_{\mathbb{R}^d} v(\mathbf{x}-\mathbf{y})\, d\mathbf{y}}$$

Here we did not assume the kernel v to be normalized.

A well-known disadvantage of linear filtering is that not only the noise is reduced but also that the image structure is smoothed. Consider the example of an image showing a detail of a scanned text (Fig. 1). The black lines are small and the grey values from two distributions will be mixed for all but the smallest neighborhood sizes. This will lead to to the weighted mean of the grey value of the text and the grey value of the white paper.

The bilateral filter prevents the mixing of two grey value distributions by introducing a tonal weight. The tonal

1051-4651/02 $17.00 (c) 2002 IEEE

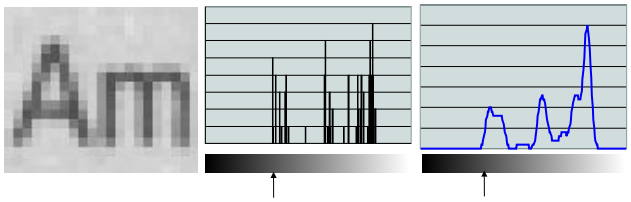

Figure 1. Noise text image patch. In a. a noisy text image patch is shown. In b. the histogram of the image patch and in c. a smoothed version of the histogram is shown.

weight depends on the tonal distance from the center point. Let w be the tonal kernel that maps tonal distance to tonal weight (just like the spatial kernel maps spatial distance to spatial weight), then the bilateral filter is defined by:

$$\frac{\int_{\mathbb{R}^d} f(\mathbf{y})\, v(\mathbf{x}-\mathbf{y})\, w(f(\mathbf{x})-f(\mathbf{y}))\, d\mathbf{y}}{\int_{\mathbb{R}^d} v(\mathbf{x}-\mathbf{y})\, w(f(\mathbf{x})-f(\mathbf{y}))\, d\mathbf{y}} \qquad (1)$$

Note that the value $f(\mathbf{y})$ in the neighborhood of the central point $\mathbf{x}$ is compared with the value $f(\mathbf{x})$ to calculate the tonal weight $w(f(\mathbf{x}) - f(\mathbf{y}))$. This choice assumes that the central value $f(\mathbf{x})$ is more or less noise free. It is a questionable assumption given the fact that we are building a noise suppression filter and therefore we should not take the central value as a good estimate for the 'real' value.

Therefore we allow for a second 'input image' g that is used as the reference tonal value in calculating the tonal weight. The assumption is that the value $g(\mathbf{x})$ provides a better estimate of the 'real' value then the value $f(\mathbf{x})$. This leads to the following definition of the *spatial-tonal normalized convolution* (henceforth abbreviated as the STN convolution).

Definition 1 (Spatial-Tonal Normalized Convolution) *The spatial-tonal normalized convolution of an image pair* $[f, g]$ *using the kernel pair* $[v, w]$ *is given by:*

$$[f,g] \circledast [v,w](\mathbf{x}) = \frac{\int_{\mathbb{R}^d} f(\mathbf{y})\; v(\mathbf{x}-\mathbf{y})\; w(g(\mathbf{x})-f(\mathbf{y}))\; d\mathbf{y}}{\int_{\mathbb{R}^d} v(\mathbf{x}-\mathbf{y})\; w(g(\mathbf{x})-f(\mathbf{y}))\; d\mathbf{y}}$$

The STN convolution is easily implemented and shows some remarkable properties in practical applications. See Fig. 2 for some examples of the STN convolution.

In all examples in this paper we will use $[v, w] = [G^s, G^t]$ with G^s the spatial Gaussian kernel at scale s and G^t the tonal Gaussian kernel at scale t.

We may observe that: (i) for $[f, g] = [f, f]$ the STN convolution equals the bilateral filter, (ii) for and $[v, w] = [v, 1]$ (i.e. G^t for $t \to \infty$) we take all grey tones equally important and then the STN convolution reduces to a classical convolution $f * v$ and (iii) small details, like lines and corners, are much better preserved compared with the classical (spatial) Gaussian convolution. The spatial scale is of little influence on these properties.

Figure 2. Spatial-tonal normalized convolution. On the left images corrupted with noise and on the right the result of the STN convolution.

3. Robust Estimation of Local Image Structure

In a zero order approximation we assume that locally an image is constant, i.e. $f(\mathbf{x}+\mathbf{y}) = f_0$ for small y. Due to noise we do not trust f_0 to be equal to $f(\mathbf{x})$, instead we will estimate that value from all neighboring values. A linear least squares estimate selects the $f_0(\mathbf{x})$ that minimizes the error:

$$\epsilon(f_0(\mathbf{x})) = \int_{\mathbb{R}^d} (f(\mathbf{y}) - f_0(\mathbf{x}))^2 G^s(\mathbf{x}-\mathbf{y}) d\mathbf{y}$$

Here we have used a Gaussian spatial 'soft aperture' G^s to select a local neighborhood. Differentiating the above quadratic error term ϵ with respect to f_0 and solving for $\partial\epsilon/\partial f_0 = 0$ we obtain:

$$\begin{aligned} f_0(\mathbf{x}) &= \frac{\int_{\mathbb{R}^d} f(\mathbf{y})\, G^s(\mathbf{x}-\mathbf{y})\, d\mathbf{y}}{\int_{\mathbb{R}^d} G^s(\mathbf{x}-\mathbf{y})\, d\mathbf{y}} \\ &= \int_{\mathbb{R}^d} f(\mathbf{y})\, G^s(\mathbf{x}-\mathbf{y})\, d\mathbf{y} = (f * G^s)(\mathbf{x}), \end{aligned}$$

i.e. the Gaussian convolution provides a least squares estimate of the zero order local image structure.

In case the image is corrupted with noise that is not normally distributed the LSQ estimate is known not to be an optimal estimate. This certainly is true for the situation that we are not dealing with noise but with the situation that the local neighborhood contains values from more then one distribution.

The constant image patch model then is not sufficient anymore. A principled solution then would be to model the local image patch as a combination of (say) two constant regions. In that case not only the value per region has to be estimated but also the geometry that separates one region from the other. A well-known example of a method that follows this scheme is the Hueckel edge detector [6].

Here we follow a less principled route. We adhere to the one-distribution constant image patch model but we allow some values in the local neighborhood to be identified as 'outliers'. To that end we will use robust estimation techniques. Instead of using a quadratic error norm that emphasizes large errors, a robust error norm will be used that does not take observations far from the values predicted by the model into account. We will use the following 'Gaussian' error norm:

$$\rho^t(p) = 1 - \exp\left(-\frac{p^2}{2t^2}\right) \tag{2}$$

that leads to the following robust error measure:

$$\epsilon(f_0(\mathbf{x})) = \int_{\mathbb{R}^d} \rho^t(f(\mathbf{y}) - f_0(\mathbf{x}))G^s(\mathbf{x}-\mathbf{y})d\mathbf{y}$$

Differentiating ϵ with respect to f_0 and solving for $d\,\epsilon/d\,f_0 = 0$ leads to:

$$\int_{\mathbb{R}^d} \phi^t(f(\mathbf{y}) - f_0(\mathbf{x}))G^s(\mathbf{x}-\mathbf{y})d\mathbf{y} = 0$$

where ϕ^t is the derivative of the robust error norm. For the Gaussian error norm in Eq. (2) we have:

$$\phi^t(p) = \frac{d\,\rho^t}{d\,p} = \frac{p}{t^2}\exp\left(-\frac{p^2}{2t^2}\right)$$

Substituting this in the equation $d\,\epsilon/d\,f_0 = 0$ results in:

$$f_0(\mathbf{x}) = \frac{\int_{\mathbb{R}^d} f(\mathbf{y})\exp\left(-\frac{(f(\mathbf{y})-f_0(\mathbf{x}))^2}{2t^2}\right)G^s(\mathbf{x}-\mathbf{y})d\mathbf{y}}{\int_{\mathbb{R}^d}\exp\left(-\frac{(f(\mathbf{y})-f_0(\mathbf{x}))^2}{2t^2}\right)G^s(\mathbf{x}-\mathbf{y})d\mathbf{y}} \tag{3}$$

The above *implicit* expression for $f_0(\mathbf{x})$ is of the form $f_0 = F(f_0)$ and can be solved using *fixed point iteration* (also known as *functional iteration*).

Given an initial estimate f^0 of the zero order local structure we may iterate $f^{i+1} = F(f^i)$ until stability to find f_0. An obvious choice for the starting value of the iteration is the value $f(\mathbf{x})$. This leads to the following theorem:

Theorem 2 ***(Robust Estimation of Zero Order Local Image Structure)*** *A robust estimator of the zero order local image structure is obtained as the asymptotic result of the the following iteration of the STN convolution:*

$$f^0 = f, \qquad f^{i+1} = [f, f^i] \circledast [G^s, G^t]$$

There are two important observations to make. The tonal scale in the STN convolution turns out to be the scale of a robust error norm. Secondly, in the iteration of the STN convolution only its second argument is changed. The spatial aperture thus is constant in the iteration process.

Robust estimation of local image structure is not new in the image processing context. Besl et al. [1] in 1989 describe the principles. What is new in our method is the use of Gaussian apertures (both spatial and tonal apertures) and the connection to mean-shift analysis and the spatially local histograms (i.e. a tonal density framework). The generalization to higher order local image structure within the same framework of STN convolutions will be reported in a forthcoming paper.

4. Local-Mode Estimation

Consider again Fig. 1. The grey value of the central point is marked with an arrow in the histogram in the middle and the smoothed histogram on the right. It is evident that the local mode closest to the grey value of the central point is a far better estimate of the 'true' grey value then the average of all grey values.

In this section we will show that iterating the STN convolution does just that: find the tonal local mode in a smoothed (spatial local) image histogram.

First we consider the STN convolution for infinite spatial scale, i.e. $v = 1$:

$$[f, g] \circledast [1, w](\mathbf{x}) = \frac{\int_{\mathbb{R}^d} f(\mathbf{y})\; w(g(\mathbf{x}) - f(\mathbf{y}))\; d\mathbf{y}}{\int_{\mathbb{R}^d} w(g(\mathbf{x}) - f(\mathbf{y}))\; d\mathbf{y}}$$

Instead of integrating over the spatial domain we may integrate over the codomain (the grey value range):

$$[f, g] \circledast [1, w](\mathbf{x}) = \frac{\int_{\mathbb{R}} p\; h_f(p)\; w(g(\mathbf{x}) - p)\; dp}{\int_{\mathbb{R}^d} h_f(p)\; w(g(\mathbf{x}) - p)\; dp}$$

where $h_f(p)dp$ is the area of $\mathbb{R}^d$ with grey values in the range from p to $p + dp$. I.e. h_f is the image *histogram*. We can rewrite the above expression:

$$[f, g] \circledast [1, w](\mathbf{x}) \;=\; g(\mathbf{x}) - \frac{\int_{\mathbb{R}} q\; h_f(g(\mathbf{x}) - q))\; w(q)\; dq}{\int_{\mathbb{R}} h_f(g(\mathbf{x}) - q))\; w(q)\; dq}$$

For $w = G^t$ we have $qG^t(q) = -t^2\partial_q G^t(q)$ and we obtain:

$$\begin{aligned}[f, g] \circledast [1, G^t](\mathbf{x}) &= g(\mathbf{x}) + t^2\frac{\partial(h_f * G^t)(g(\mathbf{x}))}{(h_f * G^t)(g(\mathbf{x}))} \\ &= g(\mathbf{x}) + t^2(\partial \log h_f * G^t)(g(\mathbf{x}))\end{aligned}$$

We thus see that the above STN convolution implements a gradient ascent in the smoothed histogram $h_f * G^t$.

Theorem 3 ***(Spatial global, tonal local mode finding)*** *The tonal local mode in the spatial global image histogram $h_f * G^t$ is found by iterating the STN convolution:*

$$f^0 = f, \qquad f^{i+1} = [f, f^i] \circledast [1, G^t]$$

until stability.

Because $v = 1$, the histogram is equal for all positions in the image, it is only the initial grey value in the iteration process that is dependent on the position (and with that the result of the iteration process). Tomasi et al. [8] already showed that the bilateral filter with a spatial scale that encompasses the entire image is equivalent with a histogram transformation. The above theorem shows that their result is the first iteration in a local mode finding process.

The image operator defined in theorem 3 acts as an image segmentation operator where the tonal domain is segmented in regions each containing one local mode in the (smoothed) image histogram. All the tonal values in such a region are replaces with the tonal value of the local mode in the histogram. Essentially this operator thus performs a watershed segmentation of the smoothed image histogram.

Instead of looking at the spatial global histogram we can look at spatial local histograms. Again we use a Gaussian 'soft aperture' G^s. Let $h_f(\mathbf{x})$ be the spatial local image histogram, then we have:

Theorem 4 ***(Spatial local, tonal local mode finding)*** *The tonal local mode in the spatial local image histogram $h_f(\mathbf{x}) * G^t$ is found by iterating the STN convolution:*

$$f^0 = f, \qquad f^{i+1} = [f, f^i] \circledast [G^s, G^t]$$

until stability.

This theorem thus shows that robust estimation of zero order local image structure is equivalent with finding the tonal local mode in the smoothed spatial local image histogram.

5. Conclusions

We have shown that the spatial-tonal normalized convolution is a generalization of the bilateral filter. The STN convolution when iterated until stability proves to be equivalent to robust estimation of zero order local image structure and to local mode finding in the spatial local image histograms.

In this paper we assumed scalar images. However the theory is easily generalized to color images. In the definition of the STN convolution the images then become vector valued. The weight functions v and w remain scalar functions. Replacing $f(\mathbf{x}) - g(\mathbf{x})$ with $\|f(\mathbf{x}) - g(\mathbf{x})\|$ then leads to an expression for the STN convolution that is valid for color images as well. In fact all examples shown in the paper are color images.

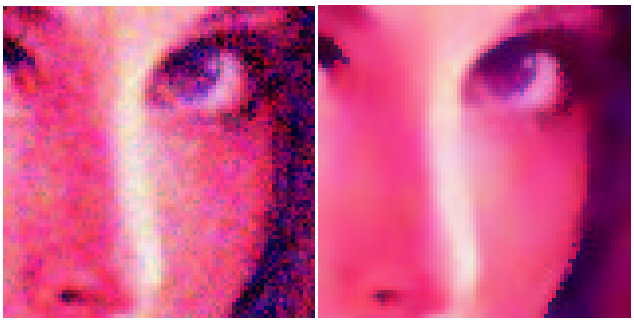

Figure 3. Iterating the spatial-tonal normalized convolution. On the left images corrupted with noise and on the right the result of the iterated STN convolution (i.e. finding the local mode, i.e. robust estimation of the zero order image structure).

The interpretation in terms of robust estimation of local image structure makes it feasible to generalize the framework to robust estimation of higher order local image structure. This will be reported in future work.

References

[1] P. Besl, J. Birch, and L. Watson. Robust window operators. *Machine Vision and Applications*, 2:179–191, 1989.

[2] M. Black, G. Sapiro, D. Marimont, and D. Heeger. Robust anisotropic diffusion. *IEEE Trans. Image Processing*, 7(3):421–432, 1998.

[3] Y. Cheng. Mean shift, mode seeking and clustering. *IEEE PAMI*, 17(8), 1995.

[4] D. Comaniciu and P. Meer. Mean shift analysis and applications. In *ICCV (2)*, pages 1197–1203, 1999.

[5] L. Grifin. Scale-imprecision space. *Image and Vision Computing*, 15:369–398, 1997.

[6] M. Heuckel. An operator which locates edges in digital pictures. *J. Association for computing machinery*, 18:113–125, 1971.

[7] J. Koenderink and A. van Doorn. The structure of locally orderless images. *Int. Journal of Computer Vision*, 31(2/3):159–168, 1999.

[8] C. Tomasi and R. Manduchi. Bilateral filtering for gray and color images. In *ICCV-98*, 1998.

[9] J. van de Weijer and R. van den Boomgaard. Local mode filtering. In *CVPR01*, pages II:428–43, 2001.