Adaptive sharpening of multimodal distributions

Freddie Åström, Michael Felsberg and Hanno Scharr

Adaptive Sharpening of Multimodal Distributions

Freddie Åström*
*Heidelberg Collaboratory for Image Processing
Heidelberg University
Germany

Hanno Scharr
IBG-2: Plant Sciences
Forschungszentrum Jülich
Germany

Michael Felsberg
*Computer Vision Laboratory
Linköping University
Sweden

***Abstract*—In this work we derive a novel framework rendering measured distributions into approximated distributions of their mean. This is achieved by exploiting constraints imposed by the Gauss-Markov theorem from estimation theory, being valid for mono-modal Gaussian distributions. It formulates the relation between the variance of measured samples and the so-called standard error, being the standard deviation of their mean. However, multi-modal distributions are present in numerous image processing scenarios, e.g. local gray value or color distributions at object edges, or orientation or displacement distributions at occlusion boundaries in motion estimation or stereo. Our method not only aims at estimating the modes of these distributions together with their standard error, but at describing the whole multi-modal distribution. We utilize the method of channel representation, a kind of soft histogram also known as population codes, to represent distributions in a non-parametric, generic fashion. Here we apply the proposed scheme to general mono- and multimodal Gaussian distributions to illustrate its effectiveness and compliance with the Gauss-Markov theorem.**

I. Introduction

We propose and investigate a method allowing to approximate the distribution of the mean from distributions of input samples, i.e. population distributions. It is applicable in image processing and computer vision methods allowing to derive population distributions explicitly. The key ingredient we focus on here is a special sharpening process converting the 'input' distribution into the 'output' distribution of the mean.

As example class of suitable image processing methods we consider smoothing methods, but other methods may be considered as well. Smoothing methods can be formulated as

$$\hat{\mathbf{g}} = \mathbf{A}\mathbf{f}\,, \tag{1}$$

where $\mathbf{f}$ is an input image, $\mathbf{A}$ is some linear or non-linear, non-negative filter, and $\hat{\mathbf{g}}$ is the desired filter output. Non-negative means that if $\mathbf{f}$ is an M-pixel image represented as a vector $f \in \mathbb{R}^M$, then $\mathbf{A}$ is a $M \times M$-matrix $\mathbf{A} \in \mathbb{R}^{M\times M}$ with non-negative entries, only. Examples for such adapted filtering are not only iterated or non-iterated local robust smoothing methods [1], [2], but also diffusion-like regularization schemes [3], [4], [5] often fulfill the non-negativity constraint. Clearly, $\mathbf{f}$ may be a gray value, color, or spectral image, and $\mathbf{A}$ may also include additional value or color transformations or projections. Here, we consider $\mathbf{g}$ being a scalar-valued image and explain how to derive population distributions.

For simplicity let us use image denoising as illustration example, where from a noisy, observed input image $\mathbf{f}$ a noise-reduced output image $\hat{\mathbf{g}}$ shall be calculated using prior knowledge on noise-free images (as e.g. in [4], [6]). In such estimation approaches the desired result is given as the location $\hat{\mathbf{g}}$ of the maximum of a suitably constructed posterior distribution $p(\mathbf{g}|\mathbf{f})$, and $\hat{\mathbf{g}}$ is called 'maximum a posteriori (MAP) estimate'. Such estimates $\hat{\mathbf{g}}$ are 'point' estimates, as only a single value (or point, here a single image $\hat{\mathbf{g}} \in \mathbb{R}^M$) is derived from the underlying distribution.

The variance of such a point estimate can then be derived using methods like the Cramer-Rao lower bound [7] by essentially estimating the curvature of the probability distribution p at the maximum location $\hat{\mathbf{g}}$, more exactly at the distribution mode selected by the estimation scheme. When interested in more than a point estimate and its variance, e.g. in the number of modes in the distribution, Monte-Carlo methods like Gibbs sampling [8] or bootstrapping [9] can help. Bootstrapping samples the distribution of $\hat{\mathbf{g}}$, by running the estimation process (i.e. here the denoising algorithm) on different samples of the input distribution $\mathbf{f}_1, ..., \mathbf{f}_N$. In our image denoising example, one may have N images showing the exact same static scene (represented by the noise-free image $\mathbf{f}_0$) but with different noise realizations $\eta_1, ...\eta_N$ such that $\mathbf{f}_i = \mathbf{f}_0 + \eta_i$. The N realizations of $\hat{\mathbf{g}}$ are then samples from the sought for distribution. However, using a reasonably small N, samples may not suffice to well represent the M-dimensional distribution $p(\hat{\mathbf{g}})$. However they may suffice to derive M marginal distributions, one at each pixel location $\mathbf{x}$, i.e. an independent distribution for each $\hat{\mathbf{g}}(\mathbf{x})$, e.g. represented as a 1d histogram. Even though simple and effective, in most application scenarios one does not have multiple images of the same scene and, obviously, computational complexity is N times the complexity of the underlying estimation algorithm.

We propose a method approximating the pixel-wise marginal distributions of the mean working with a single input image, only. Any non-parametric soft-histogram method can be used to represent the distributions. We use a channel encoded image for this (explained in Section II), i.e. each pixel contains a special soft histogram called 'channel representation' [10], [11]. Smoothing the channel encoded image yields population distributions for each pixel location (cmp. Section III, and see [2], [12]). Each population distribution is then sharpened with our proposed process introduced in Section V. For mono-modal Gaussian distribution this 'sharpening' is available in closed form via the Gauss-Markov (GM) theorem (see Section IV).

Sharpening distributions is not new. A straightforward approach to reduce the variance of a value distribution is by squaring and normalizing the density [13]. Performing this on channel representations has the drawback that density mode locations are biased towards channel centers. An approach to reduce the bias is to reconstruct a continuous representation

[14]. Despite its simplicity, squaring and normalizing the signal does not only sharpen modes, but e.g. if the distribution is bi-modal the mode with larger amplitude will suppress the smaller mode, as will be illustrated by our numerical experiments. An alternative to simple squaring is to iteratively increase the channel resolution similar to a scale-pyramid [15]. However, the end result depends on the number of channels, number of scales in the scale-pyramid and finally the number of iterations. In both of these approaches, there is no guarantee that the density after sharpening will adhere to initial assumptions.

Own contribution: We present the first sharpening method respecting the GM theorem, i.e. we derive a method for reducing variances of all modes in a distribution in the sense of robust statistics. An iterative scheme is proposed establishing a sharpening process. Our main contribution is to use the GM theorem to statistically motivate a framework reducing the variance of an estimated multimodal density.

II. Channel encoded images

The channel representation is a special soft histogram. Like storing a value u in a histogram means setting the respective bin weight to 1, encoding a value u in channels means setting (several) channel weights w_i according to

$$w_i(u) = B(u - \hat{u}_i) \quad i \in \{1, \dots, C\} \tag{2}$$

where $\hat{u}_i$ denotes equidistantly spaced channel centers, and B is a basis function with compact support. B is selected such that $\sum_{i=1}^{C} w_i = 1$, when encoding a single value. This ensures that encoding multiple values in the same channel representation can be done by summing their respective weights, as with normal histograms. For simplicity, throughout this work, B is assumed to be a B-spline of second order with compact support $[-3/2, 3/2]$, but other selections are possible. As an example Figure 2 shows an intensity value $u = 2$ corresponding to non-zero weights given by the basis functions located at $\hat{u}_i \in \{1, 2, 3\}$.

Reconstructing a single observed value u from the channel weights w_i is done by a linear combination of the C channels

$$\tilde{u} = \frac{\sum_{i=1}^{C} K_i w_i \hat{u}_i}{\sum_{i=1}^{C} K_i w_i}, \tag{3}$$

where $\mathbf{K} = [K_1, \dots, K_C]^T$ denotes a decoding window. If $K_i = 1, \forall i$ then the decoding describes a one-to-one mapping in relation to the encoded and decoded value, i.e. $\tilde{u} = u$. For multiple encoded values, their mean is reconstructed. However in practice, when several values have been encoded in the same channel representation, $\mathbf{K}$ is chosen to be a local decoding window, centered at a local maximum, reconstructing a local weighted mean, i.e. the local center of mass. This results in a robust decoding scheme.

A channel encoded image with M pixels consists of M channel representations, one at every pixel location. Thus weights of each channel i form M-pixel images $\mathbf{w}_i \in \mathbb{R}^M$.

III. From noisy image to population distributions

An input image $\mathbf{f}(\mathbf{x})$ is represented as an channel encoded image with weights $\mathbf{w}_i \in \mathbb{R}^M$. The population distribution for each pixel is then derived is two steps. First matrix $\mathbf{A}$ from (1) is normalized row-wise such that the maximum value in each each row is 1, i.e. $\tilde{\mathbf{A}} = \mathbf{D}\mathbf{A}$, where $\mathbf{D}$ is a diagonal $M \times M$-matrix with $D_{ii} = 1/\max_j A_{ij}$. This corresponds to each locally applied smoothing filter being normalized to maximum 1. The sum over a filters' values, i.e. the sum over a row of $\mathbf{A}$ corresponding to location $\mathbf{x}$, then gives a lower bound to the number of samples (pixel values) used to calculate $\hat{\mathbf{g}}(\mathbf{x})$. In a second step $\tilde{\mathbf{A}}$ is applied to each channel, $\tilde{\mathbf{w}}_i(\mathbf{x}) = \tilde{\mathbf{A}}\mathbf{w}_i(\mathbf{x})$. Please note, that $\tilde{\mathbf{A}}$ is fixed by (1) and does not depend on $\mathbf{w}_i$. The sum over all channel weights $\sum_{i=1}^{C} \mathbf{w}_i(\mathbf{x})$ at pixel location $\mathbf{x}$ is then a lower bound to the number N of encoded samples.

IV. Minimizing variances in Gaussian processes

The aim of our method is to sharpen each of the above derived pixel-wise distributions individually. For now, let us assume a mono-modal Gaussian distribution (Figure 1 (a), blue line), i.e. let g_i be a given observation for $i = 1, \dots, N$ where N is the number of observations and $\mathbf{g} = (g_1, \dots, g_N)$. Let η be additive normal distributed Gaussian noise with parameters of zero mean and variance σ^2. Then the given observation can be described by $g_i = g_0 + \eta$ or in matrix notation

$$\mathbf{g} = \mathbf{H}g_0 + \eta, \quad \mathbf{H} = [1, \dots, 1] \in \mathbb{R}^N, \tag{4}$$

which is a special case of the so-called *general linear model* [16]. The unknown parameter to be estimated is denoted by g_0 whereas $\mathbf{H}$ describes the number of measurements N. By increasing the number N of measurements g_i, the variance of the distribution of the estimated mean $\hat{g}$ will be reduced. This relation is clarified by the Gauss-Markov (GM) theorem, which states that the covariance $\mathbf{C}_{\hat{g}}$ of the estimator $\hat{g}$ of g_0 decreases quantitatively with the number of measurements used

$$\mathbf{C}_{\hat{g}} = (\mathbf{H}^T \mathbf{C}^{-1} \mathbf{H})^{-1}, \tag{5}$$

or here, as $\mathbf{C}$ is diagonal with $C_{ii} = \sigma^2$, and $\mathbf{H}$ as in (4), we get $\sigma_{\hat{g}}^2 = \frac{1}{N}\sigma^2$ for an arithmetic mean. When $\hat{g}$ is estimated as a weighted mean $\hat{g} = \sum_i \alpha_i g_i$, then

$$\sigma_{\hat{g}}^2 = \sigma^2 \sum_i \alpha_i^2. \tag{6}$$

The distribution of the mean $\hat{g}$ thus generally has smaller variance and by this is a sharpened version of the input distribution, cmp. Figure 1 (a), red line. The depicted sharpened density is optimal in the sense of GM. Because of the limited amount of data the resulting density will not get sharper. Thus, when implementing sharpening as a process, it needs to stop suitably.

We observe, that the input Gaussian distribution (Figure 1 (a), blue line), $u \in N(\mu, \sigma)$, can be transformed into the output distribution (Figure 1 (a), red line), $\hat{u} \in N(\mu, \sigma_{\hat{u}})$, according to a linear transform $\hat{u} = \frac{\sigma_{\hat{u}}}{\sigma}(u - \mu) + \mu$ or

$$\hat{u} = s \cdot (u - \mu) + \mu \tag{7}$$

with slope $s = (\sum_i \alpha_i^2)^{\frac{1}{2}}$. In Figure 1 (c) this transform is depicted for $\mu = 100$, $s \approx 1/5$. The ideal transformation given by the GM Theorem is plotted in green and the approximation by our method in red. Transforming values in a channel representation means moving weights from one channel to another. Thus sharpening a mode means moving *weights belonging to that mode* towards the mode (i.e. its center). For a channel-represented multi-modal distribution we need to (a) find out

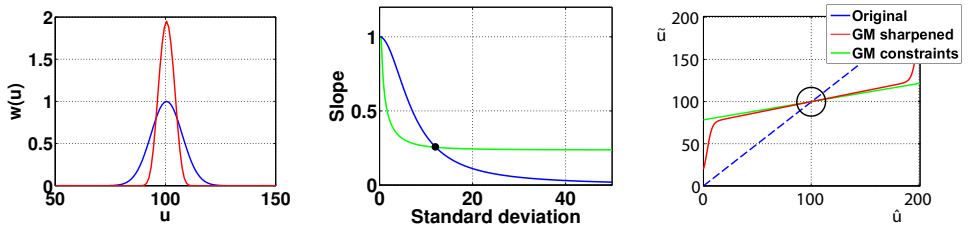

Fig. 1. Figure (a) shows one mode example of sharpening using the GM Theorem. In (b) we show the intersection point which gives the optimal kernelsize (see section V for details). In (c) we show the transition of the weights from the original to the new (sharpened) distribution.

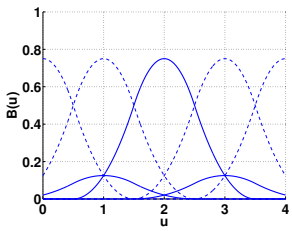

Fig. 2. Examples of 5 equidistant B-spline functions and their corresponding activation to one image sample drawn in thick lines.

which channel belongs to which mode (i.e. the local μ in (7)), and (b) how many samples belong to that mode, i.e. the local slope s. To do so, we set up an iterative process.

V. A SCHEME FOR SHARPENING MULTIMODAL DENSITIES

To obtain a sharpening process for an arbitrary distribution we perform for every channel position $\hat{u}_i$ a local decoding, yielding for every channel $i \in \{1, ..., C\}$ a value $\tilde{u}_i$. This decoding is followed by encoding the values $\tilde{u}_i$, yielding new channel weights. Compared to $\hat{u}_i$, as $\tilde{u}_i$ is a local center of mass, $\tilde{u}_i$ is *shifted towards the local mode*, i.e. 'up-hill' in the distribution. Local decoding and encoding are iterated. Local decoding means that each $\tilde{u}_i$ is obtained by decoding with a different kernel $\mathbf{K}_i$, centered at channel position $\hat{u}_i$, according to (3). By this procedure we get a sharpening process which can be steered at every single channel position by the size (and shape) of the decoding kernel (or window) $\mathbf{K}_i$. This sharpening process is ad hoc. However, the advantage is the possibility to steer the sharpening by selecting $\mathbf{K}_i$. In this work, kernels $\mathbf{K}_i$ are assumed to be truncated Gaussian kernels. Their variance $\sigma_{\mathbf{K}_i}$ is derived such that the process conforms with the GM theorem, as shown next.

To derive the estimation of optimal kernel sizes, let us for a moment assume to use the same decoding kernel $\mathbf{K}$ for all channels. The larger $\mathbf{K}$, the lower is slope $\tilde{s}$ of the function transforming $\hat{u}_i$ into $\tilde{u}_i$. In the extreme case $\sigma_{\mathbf{K}} \to 0$ values $\hat{u}_i$ are reproduced $\tilde{u}_i = \hat{u}_i$, i.e. slope $\tilde{s} = 1$. This is consistent with (6), as then only one of the values K_i is different from 0 and thus, according to the normalization in (3), the corresponding $\alpha_i = 1$, all others are 0. This means $\sigma_{\hat{g}}^2 = \sigma^2$. In the other extreme case $\sigma_{\mathbf{K}} \to \infty$ we get $\tilde{u}_i$ as the mean over the whole input distribution, no matter where we center $\mathbf{K}$, i.e. slope $\tilde{s} = 0$. This is not consistent with (6), as then from (3) we get

$$\tilde{u} = \frac{\sum_{i=1}^{C} K_i w_i \hat{u}_i}{\sum_{i=1}^{C} K_i w_i} \approx \frac{\sum_{i=1}^{C} w_i \hat{u}_i}{\sum_{i=1}^{C} w_i}, \tag{8}$$

because all values $K_i \approx$ const. in the interval where w_i are non-zero. Consequently $\alpha_i = w_i / \sum_i w_i$. As $\sum_i w_i \leq N \leq M$ we get for slope $s = \sum_i \alpha_i^2$ that $1/N \leq s \leq 1$.

To find an optimal kernel for a channel i we calculate slopes $\tilde{s}$ and s (blue and green lines in Figure 1 (b), respectively) as a function of kernel variance $\sigma_{\mathbf{K}_i}$. For this we calculate $\tilde{u}_i$ using (3) and get $\tilde{s}$ using finite differences. GM-conform slope s we calculate combining (3), (6) and its definition from (7)

$$s = \left(\sum_{i=1}^{C} \left(\frac{K_i w_i}{\sum_{j=1}^{C} K_j w_j} \right)^2 \right)^{\frac{1}{2}} \tag{9}$$

Starting from $\sigma_{\mathbf{K}_i} \gtrsim 0$ with increasing $\sigma_{\mathbf{K}_i}$ we select the first kernel where either $\tilde{s} = s$ or where $\tilde{s}$ has a first minimum. A minimum indicates that kernel $\mathbf{K}_i$ reaches into a second mode.

Figure 1 (a) shows a unimodal distribution which has been sharpened with the proposed method. In (c) the ideal slope given by the GM Theorem can be seen in green. The corresponding rate of change of the given distribution at the current channel position can be seen in blue in (b). The position where the optimal kernel size is determined is chosen to be $\hat{u}_i = 100$. The dot in (b) indicates the chosen standard deviation which fits the GM slope. By looking at the given signal in (a) we observe that the chosen standard deviation of around 13 (and therefore the kernel size) mainly includes all information of the signal for the sharpening process.

VI. ITERATIVE SHARPENING OF MULTIPLE MODES

The presence of multiple modes requires the consideration of mixed distributions when modes overlap. Here we consider a bimodal distribution. Figure 3 (a) shows a distribution encoded in 21 channels (black crosses) and the corresponding continuous reconstruction which is done using the maximum entropy method described in [14]. Other reconstruction methods are possible but this is outside the scope of this paper. In the same figure the selected kernel used for sharpening is shown at the channel position of around 85, the standard deviation (approx. 7) of this kernel was obtained from the point indicated in (b). We selected this location since it is the local minimum of the slope of the corresponding sharpened distribution. Because of the shape of the given intensity distribution the optimal change of weights cannot be obtained in a single step but in several iterations. In terms of relation to the distribution, it means if the kernel variance is increased then mass from the neighboring mode will be collected which is an unwanted effect. Figure (c) shows the relative change of the weights for the selected kernel. At the indicated position it is clear that the slope of the sharpened distribution conforms with the slope of the GM theorem. As expected this only holds in a neighborhood around the current position.

Figures 3 (d)–(f) show the sharpening result after one, two and three iterations. We compared sharpening by the novel GM

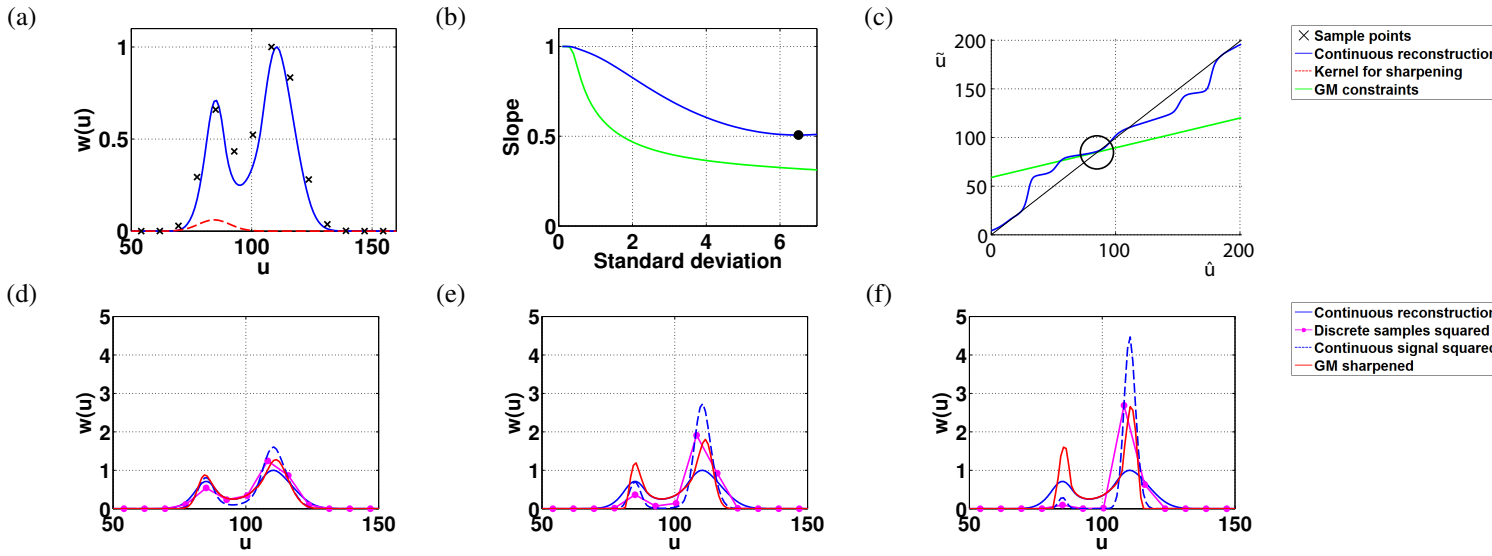

Fig. 3. In (a) we show samples of an example distribution, its continuous reconstruction (blue) and the selected kernel used for sharpening (red). (b) shows the slope given by the GM theorem (green) and the sharpened density (blue). (c) shows the transformation of the distribution derived using optimal kernels. (d)-(f) illustrate the iterative behavior of the proposed scheme compared to squaring and normalization of the discrete samples and the continuous reconstruction (see text). Note that in (f) it is clearly visible that simple squaring of the distribution will suppress the smaller mode. A drawback not present in our framework.

formulation and simple squaring of the distribution, which we call the *standard method*. The standard method was applied to both the continuous representation of the density as well as its discrete sample points. It is clear that the approach yields a strong bias towards the channel center when considering the discrete samples. This is visible in the mode with higher amplitude. For the continuous representation the bias is not severe. However as the number of iterations increases the relative mass distribution between the two modes changes such that the mode with larger amplitude suppresses the smaller mode. We observe that the GM sharpening process does not suffer from the mentioned drawbacks.

VII. Conclusion

A method for reducing the variance of single modes in multimodal distributions has been presented, which is in agreement with the Gauss-Markov theorem. The derived sharpened distribution is therefore an approximation of the distribution of the local mean of the corresponding mode. We have shown in numerical experiments the superior performance of our approach compared to sharpening by squaring, and that our proposed scheme indeed yields an optimal sharpening of multimodal distributions.

Acknowledgment

This research has received funding from the Swedish Foundation for Strategic Research through the grant VPS and from Swedish Research Council through grants for the projects energy models for computational cameras (EMC^2), Visualization-adaptive Iterative Denoising of Images (VIDI), all within the Linnaeus environment CADICS and the excellence network ELLIIT. Support by the German Science Foundation and the Research Training Group (GRK 1653) is gratefully acknowledged by the first author.

References

[1] C. Tomasi and R. Manduchi, "Bilateral filtering for gray and color images," in *ICCV '98*, 1998, pp. 839–846.

[2] M. Felsberg, P.-E. Forssen, and H. Scharr, "Channel smoothing: Efficient robust smoothing of low-level signal features," *IEEE Trans. PAMI*, vol. 28, no. 2, pp. 209–222, 2006.

[3] P. Perona and J. Malik, "Scale-space and edge detection using anisotropic diffusion," *IEEE Trans. PAMI*, vol. 12, pp. 629–639, 1990.

[4] H. Scharr, M. J. Black, and H. W. Haussecker, "Image statistics and anisotropic diffusion," in *ICCV*, 2003, pp. 840–847.

[5] S. Roth and M. J. Black, "Fields of experts," *International Journal of Computer Vision (IJCV)*, vol. 2, pp. 205–229, April 2009.

[6] D. Mumford and J. Shah, "Optimal approximations by piecewise smooth functions and associated variational problems," *Communications on Pure and Appl. Math.*, vol. 42, no. 5, pp. 577–685, 1989.

[7] S. Kay, *Fundamentals of statistical signal processing*. Englewood Cliffs, N.J: Prentice-Hall PTR, 1993.

[8] K. Krajsek, I. Dedovic, and H. Scharr, "An estimation theoretical approach to ambrosio-tortorelli image segmentation," in *DAGM*, 2011, pp. 41–50.

[9] B. Efron, "Bootstrap methods: Another look at the jackknife," *Ann. Statist.*, vol. 7, no. 1, pp. 1–26, 01 1979.

[10] P.-E. Forssén, "Low and medium level vision using channel representations," Ph.D. dissertation, Linköping University, Sweden, March 2004, dissertation No. 858, ISBN 91-7373-876-X.

[11] G. H. Granlund, "An associative perception-action structure using a localized space variant information representation," in *Proc. AFPAC*, Kiel, Germany, Sept. 2000.

[12] M. Felsberg and G. Granlund, "Anisotropic Channel Filtering," in *SCIA*, ser. LNCS, vol. 2749, 2003, pp. 755–762.

[13] S. L. Deneve, P. E., and A. Pouget, "Reading population codes: a neural implementation of ideal observers," *Nature America Inc.*, vol. 2, no. 8, pp. 740–745, 1999.

[14] E. Jonsson and M. Felsberg, "Reconstruction of probability density functions from channel representations," in *Image Analysis*, ser. LNCS. Springer Berlin Heidelberg, 2005, vol. 3540, pp. 491–500.

[15] M. Felsberg, "Incremental computation of feature hierarchies," in *Pattern Recognition, LNCS 6376*. Springer, 2010, pp. 523–532.

[16] S. M. Kay, *Fundamentals of Statistical Signal Processing, Volume 1: Estimation Theory*. Pearson Education, 1993.