Reconstruction of Probability Density Functions from Channel Representations^{*}

Erik Jonsson and Michael Felsberg

Computer Vision Laboratory,
Dept. of Electrical Engineering, Linköping University
`erijo@isy.liu.se`, `mfe@isy.liu.se`

Abstract. The channel representation allows the construction of soft histograms, where peaks can be detected with a much higher accuracy than in regular hard-binned histograms. This is critical in e.g. reducing the number of bins of generalized Hough transform methods. When applying the maximum entropy method to the channel representation, a minimum-information reconstruction of the underlying continuous probability distribution is obtained.

The maximum entropy reconstruction is compared to simpler linear methods in some simulated situations. Experimental results show that mode estimation of the maximum entropy reconstruction outperforms the linear methods in terms of quantization error and discrimination threshold. Finding the maximum entropy reconstruction is however computationally more expensive.

1 Introduction

Many methods in computer vision achieve robustness from a voting and clustering approach. Some examples are the *Hough transform* [1, 14], object recognition [10] and view matching [1, 11]. The methods all have in common that each measurement casts a vote, and that clusters in vote space define the output. In high-dimensional spaces, the simple vote histogram as used in the Hough transform is infeasible, since the number of bins will explode.

The construction of soft histograms using the *channel representation* is a practical way to reduce the number of bins required without decreasing the accuracy of the peak detection. Where previous channel decoding methods [6, 4] were designed merely to extract peaks, this paper addresses the more general problem of reconstructing a continuous probability density function from a channel vector. The theoretically sound *Maximum Entropy Method* (Sect. 3) as well as a simpler linear minimum-norm method (Sect. 4.1) is considered. From these continuous reconstructions modes can be extracted, and in Sect. 5 some properties of these modes are compared to those detected by the previous *virtual shift decoding* [4].

^{*} This work has been supported by EC Grant IST-2003-004176 COSPAL.

2 Channel Representations and Soft Histograms

In the *channel representation* [8, 5, 7], a value x is encoded into a *channel vector* $\mathbf{c}$ using the nonlinear transformation

$$\mathbf{c} = [c_1, \dots, c_N] = [K(x - \xi_1), \dots, K(x - \xi_N)] , \quad (1)$$

where K is a localized symmetric non-negative kernel function and $\xi_n, n \in [1, N]$ the *channel centers*, typically located uniformly and such that the kernels overlap. In this paper, we assume integer spacing between channels, and use the quadratic B-spline kernel [15], defined as

$$K(x) = \begin{cases} \frac{3}{4} - x^2 & 0 \leq |x| \leq \frac{1}{2} \\ \frac{1}{2}(\frac{3}{2} - |x|)^2 & \frac{1}{2} < |x| \leq \frac{3}{2} \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

Assuming that we have a number of samples of some feature, each sample can be channel encoded and the corresponding channel vectors summed or averaged. This produces a *soft histogram* - a histogram with partially overlapping and smooth bins. In a regular histogram, each sample is simply put in the closest bin, and we cannot expect to locate peaks in such histograms with greater accuracy than the original bin spacing. In a soft histogram however, it is possible to locate peaks with high accuracy at a sub-bin level. A procedure to find such peaks is referred to as a *decoding* of the soft histogram (channel vector). There are several decoding methods [6, 4], all of which are capable of perfect reconstruction of a single encoded value x .

The representation (1) can be extended into higher dimensions in a straightforward way by letting x and ξ_n be vectors, and taking the vector norm instead of the absolute value. Another option is to form the 1D channel representation in each dimension and take the outer product, giving a separable representation.

To make a connection between channel vectors and probability density functions, consider a sample set $\{x^{(m)}\}$ of size M and its encoding into channel vectors using N channels. Let $\mathbf{c} = [c_1, \dots, c_N]$ be the average of the channel representations of all samples, such that

$$c_n = \frac{1}{M} \sum_{m=1}^M K(x^{(m)} - \xi_n) . \quad (3)$$

If we assume that the samples $x^{(m)}$ are drawn from a distribution with density p , the expectation of c_n is

$$\mathbb{E}[c_n] = \int_{-\infty}^{\infty} p(x) K(x - \xi_n) dx , \quad (4)$$

showing that the elements in $\mathbf{c}$ estimate some linear features of the density function p [4]. The main focus of this paper is on how to reconstruct p from the channel coefficients c_n .

3 Maximum Entropy Method

The problem of reconstructing a continuous distribution from a finite number of feature values is clearly underdetermined, and some regularization has to be used. The natural regularizer for density functions is the *entropy*, which measures the information content in a distribution, such that the distribution with maximal entropy is the one which contains a minimum of spurious information [2].

3.1 Problem Formulation

Using the *Maximum Entropy Method* (MEM), the problem becomes the following: Find the distribution p that maximizes the (differential) entropy

$$H(p) = - \int_{-\infty}^{\infty} p(x) \log p(x) dx \quad (5)$$

under the constraints

$$\int_{-\infty}^{\infty} p(x) K(x - \xi_n) dx = c_n, \quad 1 \leq n \leq N \quad (6)$$

$$\int_{-\infty}^{\infty} p(x) dx = 1 \quad (7)$$

Using variational calculus, it can be shown that the solution is of the form [2, 9]

$$p(x) = k \exp \left(\sum_{n=1}^N \lambda_n K(x - \xi_n) \right), \quad (8)$$

where k is determined by the constraint (7). It remains now to find the set of λ_n 's which fulfills the constraints (6-7). This is a non-linear problem, and we must resort to numerical methods. In the next section, a Newton method for finding the λ_n 's is presented.

3.2 Newton Method

To make the notation more compact, we introduce a combined notation for the constraints (6 - 7) by defining feature functions f_n and residuals

$$f_n(x) = K(x - \xi_n) \quad \text{for } 1 \leq n \leq N \quad (9)$$

$$f_{N+1} \equiv 1 \quad (10)$$

$$r_n = \int_{-\infty}^{\infty} p(x) f_n(x) dx - d_n \quad \text{for } 1 \leq n \leq N+1, \quad (11)$$

where $\mathbf{d} = [c_1, \dots, c_N, 1]^T$. In this notation, (8) becomes

$$p(x) = \exp \left(\sum_{n=1}^{N+1} \lambda_n f_n(x) \right). \quad (12)$$

We can now apply a Newton method on the system of equations $\mathbf{r} = 0$. Noting that

$$\frac{dr_i}{d\lambda_j} = \int_{-\infty}^{\infty} \frac{dp(x)}{d\lambda_j} f_i(x) dx = \quad (13)$$

$$= \int_{-\infty}^{\infty} f_i(x) f_j(x) p(x) dx \quad , \quad (14)$$

the Jacobian becomes

$$\mathbf{J} = \left[\frac{dr_i}{d\lambda_j} \right]_{ij} = \left[\int_{-\infty}^{\infty} f_i(x) f_j(x) p(x) dx \right]_{ij} . \quad (15)$$

The update in the Newton method is now $\boldsymbol{\lambda} \leftarrow \boldsymbol{\lambda} + \mathbf{s}$, where the increment $\mathbf{s}$ in each step is obtained by solving the equations $\mathbf{J}\mathbf{s} = -\mathbf{r}$.

Since our feature functions f_n are localized functions with compact support, most of the time f_i and f_j will be non-zero at non-overlapping regions such that $f_i f_j \equiv 0$, making $\mathbf{J}$ band-shaped and sparse, and hence easy to invert. The main work load in this method is in the evaluation of the integrals, both for $\mathbf{J}$ and $\mathbf{r}$. These integrals are non-trivial, and must be evaluated numerically.

4 Linear Methods

Since the MEM solution is expensive to calculate, we would like to approximate it a simpler and faster approach. In Sect. 4.1, a computationally efficient linear method for density reconstruction is presented, and in Sect. 4.2, the decoding method from [4] is reviewed.

4.1 Minimum-Norm Reconstruction

If we relax the positivity constraints on $p(x)$, we can replace the maximum-entropy regularization with a *minimum norm* (MN) regularization, which permits the use of linear methods for the reconstruction. For the derivations, we consider the vector space $L_2(\mathbb{R})$ of real square-integrable functions [3], with scalar product

$$\langle f, g \rangle = \int_{-\infty}^{\infty} f(x) g(x) dx \quad (16)$$

and norm $\|f\|^2 = \langle f, f \rangle$. The minimum norm reconstruction problem is now posed as

$$p_* = \arg \min_p \|p\|^2 \quad \text{subject to } \langle p, f_n \rangle = d_n \quad \text{for } 1 \leq n \leq N+1 . \quad (17)$$

Reconstructing p from the d_n 's resembles the problem of reconstructing a function from a set of frame coefficients [12]. The reconstruction p_* of minimum norm is in $Q_1 = \text{span}\{f_1, \dots, f_{N+1}\}$, which can easiest be seen by decomposing

p_* into $p_* = q_1 + q_2$, where $q_1 \in Q_1$ and $q_2 \in Q_1^\perp$. Since $q_1 \perp q_2$, we have $\|p_*\|^2 = \|q_1\|^2 + \|q_2\|^2$. But $q_2 \perp f_n$ for all feature functions f_n , so q_2 does not affect the constraints and must be zero in order to minimize $\|p_*\|^2$. Hence $p_* = q_1 \in Q_1$, which implies that p_* can be written as

$$p_*(x) = \sum_n \alpha_n f_n(x) . \quad (18)$$

To find the set of α_n 's making p_* fulfill the constraints in (17), write

$$d_n = \langle p_*, f_n \rangle = \left\langle \sum_k \alpha_k f_k , f_n \right\rangle = \sum_k \alpha_k \langle f_k, f_n \rangle , \quad (19)$$

giving the α_n 's as a solution of a linear system of equations. In matrix notation, this system becomes

$$\Phi \alpha = \mathbf{d} , \quad (20)$$

where $\alpha = [\alpha_1, \dots, \alpha_{N+1}]^T$ and Φ is the Gram matrix

$$\Phi = [\langle f_i, f_j \rangle]_{ij} . \quad (21)$$

Note that since Φ is independent of our feature values $\mathbf{d}$, it can be calculated analytically and inverted once and for all for a specific problem class. The coefficients α can then be obtained by a single matrix multiplication.

A theoretical justification of using the minimum norm to reconstruct density functions can be given in the case where p shows just small deviations from a uniform distribution, such that $p(x)$ is defined on $[0, K]$, and $p(x) \approx K^{-1}$. In this case, we can approximate the entropy by linearizing the logarithm around K^{-1} using the first terms of the Taylor expansion:

$$H(p) = - \int_0^K p(x) \log p(x) dx \approx - \int_0^K p(x) (\log K^{-1} + K p(x)) dx = \quad (22)$$

$$= \log(K) \int_0^K p(x) dx - K \int_0^K p(x)^2 dx . \quad (23)$$

Since $\int_0^K p(x) dx = 1$, maximizing this expression is equivalent to minimizing $\int_0^K p(x)^2 dx = \|p\|^2$. This shows that the closer $p(x)$ is to being uniform, the better results should be expected from the minimum-norm approximation.

4.2 Efficient Mode Seeking

Assume that we are not interested in reconstructing the exact density function $p(x)$, but just need to locate modes of the distribution with high accuracy. An efficient method for this problem, called the *virtual shift decoding*, is introduced in [4] and briefly reviewed in this section. Let $\mathbf{c} = [c_1, \dots, c_N]$ be the averaged

channel vector of samples drawn from a distribution $p(x)$ as earlier. For simplicity, assume that the channel centers are located at the positive integers. We want to do M-estimation on $p(x)$ by finding minima of the risk function

$$\mathcal{E}(x) = (\rho * p)(x) \quad (24)$$

where ρ is some robust error norm¹. The exact shape of this error norm is not that important; what is desired is to achieve good performance in terms of e.g. low channel quantization effects [4]. Choose an error norm ρ with derivative

$$\rho'(x) = K(x - 1) - K(x + 1) , \quad (25)$$

where K is our quadratic B-spline kernel from (2). To find extrema of $\mathcal{E}(x)$, we seek zeros of the derivative

$$\mathcal{E}' = \rho' * p = (K(\cdot - 1) - K(\cdot + 1)) * p = \quad (26)$$

$$= K(\cdot - 1) * p - K(\cdot + 1) * p . \quad (27)$$

Our channel coefficients c_k can be seen as the function $K * p$, sampled at the integers. By letting $c'_n = c_{n-1} - c_{n+1}$, the new coefficients c'_n become samples of the function $\mathcal{E}'(x)$ at the integers. To find the zeros of $\mathcal{E}'$, we can interpolate these sampled function values using quadratic B-splines again. Since the interpolated $\mathcal{E}'$ can be written as a linear combination of a number of B-spline kernels (which are piecewise quadratic polynomials), the zeros of this expansion can then be found analytically.

In practise, a recursive filtering [15] or the FFT is used to find the interpolation of $\mathcal{E}'$, and the analytic solution of the zero crossings is only determined at positions where the original channel encoding has large values from the beginning, which leads to a computationally efficient method. We will not go into more detail about this, but refer to [4]. The main point is that in this mode seeking scheme, it is the derivative of some risk function which is reconstructed, and not the actual underlying probability density.

5 Experimental Results

In this section, the qualitative and quantitative behavior of the two reconstruction methods are compared. Since the MEM requires the numerical evaluation of integrals, all methods are discretized using 400 samples per unit distance. The continuous functions p and f_n are replaced by vectors, and the integrals replaced by sums. The channel centers are still assumed to be at the integers of the original continuous domain.

As a channel coefficient c_n gets closer to zero, the corresponding λ_n from the MEM tends towards $-\infty$, leading to numerical problems. To stabilize the solution in these cases, a small background DC level is introduced (ϵ -regularization).

¹ Some function which looks like a quadratic function close to zero but saturates for large argument values.

5.1 Qualitative Behavior

In Fig. 1, the qualitative behavior of the MEM and MN reconstructions are examined. The feature vector $\mathbf{d}$ was calculated for some different distributions (with the channel centers located at the integers). The two Gaussians (c-d) are reconstructed almost perfectly using MEM, but rather poorly using MN. In (b), the two rightmost peaks are close enough to influence each other, and here even the maximum entropy reconstruction fails to find the exact peak locations. For the slowly varying continuous distribution (a), both methods perform quite well.

Fig. 1. The MEM and MN reconstruction of (a) Sum of two Gaussians, (b) 4 Diracs of different weights, (c-d) Single Gaussians with different variance

5.2 Quantitative Behavior

To make a more quantitative comparison, we focused on two key properties; the *discrimination threshold* [13] and the *quantization effect* [4] of channel decoding. These properties can be measured also for the virtual shift mode seeking.

Recall that the latter decoding finds the minimum of an error function, which is equivalent to finding the maximum of the density function convolved with some kernel. To estimate a similar error function minimum from the continuous reconstructions, our estimated p should likewise be convolved with some kernel

prior to the maximum detection. Let $K_{VS} = \rho_{\max} - \rho(x)$ be the kernel implicitly used in the virtual shift decoding. For all continuous reconstruction experiments, peaks were detected from the raw reconstruction p as well as from $K * p$ (with the standard B-spline kernel K) and $K_{VS} * p$.

To measure the discrimination threshold, two values $x_0 \pm d$ were encoded. The discrimination threshold in this context is defined as the minimum value of d which gives two distinct peaks in the reconstruction. As the background DC level increases, the distribution becomes closer to uniform, and the performances of the MEM and MN methods are expected to become increasingly similar. To keep this DC level low but still avoid numerical problems, we chose a regularization level corresponding to 1% of the entire probability mass. The discrimination threshold was calculated for both reconstruction methods and the different choices of convolution kernels, and the results are summarized in Table 1 for x_0 at a channel center ($\Delta x_0 = 0$) as well as in the middle between two centers ($\Delta x_0 = 0.5$).

Table 1. Discrimination thresholds

Method		$\Delta x_0 = 0$	$\Delta x_0 = 0.5$
MEM	p	0.34	0.57
	$K * p$	0.53	0.71
	$K_{VS} * p$	1.00	1.00
MN	p	0.57	0.71
	$K * p$	0.64	0.81
	$K_{VS} * p$	0.95	1.00
Virtual Shift		1.00	1.00

With quantization effect, we mean that two distributions p differing only in shift relative to the grid of basis functions are reconstructed differently. To measure this effect, two distinct impulses of equal weight located at $x_0 \pm d$ with d below the discrimination threshold were encoded. Ideally, the peak of the reconstructed distribution would be located at x_0 , but the detected position m actually varies depending on the location relative to the grid. Figure 2 shows the difference between the the maximum of the reconstruction and the ideal peak location for varying positions and spacing of the two peaks, using the MEM and MN reconstruction. Figure 3 shows the same effect for the virtual shift decoding. Note the different scales of the plots. Also note that as the error becomes small enough, the discretization of $p(x)$ becomes apparent.

Achieving a quantization error as low as 1% of the channel distance in the best case in a very nice result, but keep in mind that this error is only evaluated for the special case of two Diracs. For smooth distributions, we expect a lower effect. It is not clear to the authors how to measure this effect in a general way, without assuming some specific distribution.

Fig. 2. The quantization effect for continuous reconstructions. Top: Maximum entropy. Bottom: Minimum norm. Left: Raw density estimate. Middle, Right: Density estimate convolved with K , K_{VS}

Fig. 3. The quantization effect for virtual shift decoding

6 Conclusions and Outlook

The maximum entropy reconstruction is theoretically appealing, since it provides a natural way of reconstructing density functions from partial information. In most applications however, the exact shape of the density function is not needed, since we are merely interested in locating modes of the distribution with high accuracy. Efficient linear methods for such mode detection can be constructed, but generally perform worse in terms of quantization error and discriminating capabilities than possible using an entropy-based method. However, as the distributions become more uniform, the performances of the approaches are expected to become increasingly similar.

In general, for convolution kernels in the mode extraction, we get better position invariance but worse discriminating capabilities. Thus, there is a trade-off between these two effects. The possibility of achieving as little quantization error as $\pm 1\%$ of the channel distance opens up for the use of channel-based

methods in high-dimensional spaces. In e.g. Hough-like ellipse detection, the vote histogram would be 5-dimensional, and keeping a small number of bins in each dimension is crucial. Unfortunately, turning the MEM reconstruction method into a practical and efficient mode seeking algorithm is nontrivial, and finding high-performance decoding methods for high-dimensional data is an issue for future research.

References

1. D.H. Ballard. Generalizing the Hough transform to detect arbitrary patterns. *Pattern Recognition*, 2(13):111–122, 1981.
2. Adam Berger, S. Della Pietra, and V. Della Pietra. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):790–799, 1996.
3. Lokenath Debnath and Piotr Mikusiński. *Introduction to Hilbert Spaces with Applications*. Academic Press, 1999.
4. M. Felsberg, P.-E. Forssén, and H. Scharr. Efficient robust smoothing of low-level signal features. Technical Report LiTH-ISY-R-2619, Dept. EE, Linköping University, SE-581 83 Linköping, Sweden, August 2004.
5. M. Felsberg and G. Granlund. Anisotropic channel filtering. In *Proc. 13th Scandinavian Conference on Image Analysis*, LNCS 2749, pages 755–762, Gothenburg, Sweden, 2003.
6. Per-Erik Forssén. *Low and Medium Level Vision using Channel Representations*. PhD thesis, Linköping University, Sweden, SE-581 83 Linköping, Sweden, March 2004. Dissertation No. 858, ISBN 91-7373-876-X.
7. Per-Erik Forssén and Gösta Granlund. Sparse feature maps in a scale hierarchy. In *AFPAC, Algebraic Frames for the Perception Action Cycle*, Kiel, Germany, September 2000.
8. G. H. Granlund. An associative perception-action structure using a localized space variant information representation. In *Proceedings of Algebraic Frames for the Perception-Action Cycle (AFPAC)*, Kiel, Germany, September 2000.
9. Simon Haykin. *Neural Networks, A Comprehensive Foundation*. Prentice Hall, second edition, 1999.
10. Björn Johansson. *Low Level Operations and Learning in Computer Vision*. PhD thesis, Linköping University, Sweden, SE-581 83 Linköping, Sweden, December 2004. Dissertation No. 912, ISBN 91-85295-93-0.
11. David G. Lowe. Object recognition from local scale-invariant features. In *CVPR'01*, 2001.
12. Stéphane Mallat. *A Wavelet Tour of Signal Processing*. Academic Press, 1998.
13. Herman P. Snippe and Jan J. Koenderink. Discrimination thresholds for channel-coded systems. *Biological Cybernetics*, 66:543–551, 1992.
14. M. Sonka, V. Hlavac, and R. Boyle. *Image Processing, Analysis, and Machine Vision*. Brooks / Cole, 1999.
15. Michael Unser. Splines: A perfect fit for signal and image processing. *IEEE Signal Processing Magazine*, 16(6):22–38, November 1999.