A BAYESIAN APPROACH TO ROBUST LOCAL FACET ESTIMATION

Robert M. Haralick

Virginia Polytechnic Institute and State University, Blacksburg,
VA 24061

C. R. Smith and G. J. Erickson (eds.),
Maximum-Entropy and Bayesian Spectral Analysis and Estimation Problems, 85–97.
© 1987 by D. Reidel Publishing Company.

1. Introduction

The facet model for image processing takes the observed pixel values to be a noisy discretized sampling of an underlying gray tone intensity surface that in each neighborhood of the image is simple. To process the image requires the estimation of this simple underlying gray tone intensity surface in each neighborhood of the image. Prewitt (1970), Haralick and Watson (1981), and Haralick (1980, 1982, 1983, 1984) all use a least squares estimation procedure. In this note we discuss a Bayesian approach to this estimation problem. The method makes full use of prior probabilities. In addition, it is robust in the sense that it is less sensitive to small numbers of pixel values that might deviate highly from the character of the other pixels in the neighborhood.

Two probability distributions define the model. The first distribution specifies the conditional probability density of observing a pixel value, given the true underlying gray tone intensity surface. The second distribution specifies the conditional probability density of observing a neighborhood having a given underlying gray tone intensity surface.

To motivate the representations we choose, and to help make clear what underlying gray tone intensity surface means, consider the following thought experiment. Suppose we have a noiseless image that is digitized to some arbitrary precision. Suppose, for the moment, we take simple underlying gray tone intensity surface to mean a constant surface in each neighborhood. Now begin moving a fixed and reasonable sized neighborhood window through the image. Most neighborhoods (probably all of them) will not have constant values. Many would be constant except for illumination shading or texture effects; those neighborhoods are nearly constant. Some have an object edge passing through; these are not constant.

The nearly constant neighborhoods can be thought of as having arisen from small perturbations of a constant neighborhood. The perturbation is due, not to sensor noise, but to the difference between the idealization of the model (perfectly constant neighborhoods) and the observed perfect reality. In this case, we take the underlying gray tone intensity surface to be a constant, the value of which is representative of the values in the observed nearly constant neighborhood.

What does it mean to determine a value that is representative of the values in the neighborhood? Does it mean an equally weighted average, for example? To answer this question, fix attention on the center pixel of the neighborhood. We expect that the neighbors of the center pixel would have a value close to the value of the center pixel. The neighbors of these neighbors, the second neighbors, would have values that could deviate more from the value of the center pixel than the first neighbors. This expectation—that the closer a pixel is to the center pixel, the less the deviation is likely to be from the center pixel—should find a way to be incorporated into the model explicitly. Under these conditions, the representative gray tone intensity of the underlying gray tone intensity surface in the neighborhood can be estimated as an unequally weighted average of the pixel values in the neighborhood, those pixels farther away from the center pixel getting less weight.

We have neglected the neighborhoods having an edge or a line passing through them. These neighborhoods do not satisfy the spirit of a model that is "constant in each neighborhood." This suggests that we need to be examining models in which the spatial distribution of gray tones in the neighborhood is more complex than constant. An appropriate model, for example, may be one in which the ideal gray tone intensity surface is a low order polynomial of row and column positions.

Now suppose that our model is that the underlying gray tone intensity surface in each neighborhood is a bivariate cubic polynomial. Again take our hypothetical noiseless perfect image and pass a neighborhood window through it. As before, there probably will be no neighborhoods that fit a cubic precisely, but this time most neighborhoods will nearly or almost nearly fit. The cubic model can represent constants, slopes, edges, and lines.

Fix attention on one of the neighborhoods. Suppose it is mostly constant, especially near its center, with a small portion of a line or edge at its boundary. Instead of thinking of the polynomial underlying gray tone intensity surface as representative, in the sense of fitting, of the entire neighborhood, think of it as containing the values of all partial derivatives of order 3 or less evaluated at the center pixel. Since the area around the center pixel is nearly constant, we should expect all partial derivatives of order 1 to order 3 to be small or zero, despite some significant disturbance at the boundary of the neighborhood and despite the fact that a least squares fit of the pixel values in the neighborhood would certainly not produce near-zero partial derivatives.

At this point we begin to uncover a few concepts about which deeper understanding is needed. The first is the difference between estimating the derivatives at the center pixel of a neighborhood and least squares fitting an entire neighborhood. The second is the notion of neighborhood size. The larger the neighborhood, the more different things are likely to happen near and around its boundary and the more we will want to ignore the values around the boundary in estimating the partial derivatives at the neighborhood center. At the same time, should the pixel values near and around the boundary of the neighborhood fit in with the spatial distribution at the center of the neighborhood, we would definitely want to have the estimation procedure utilize these neighborhood boundary pixels in a supportive way.

The conclusion we can draw from this perspective is that we can expect the underlying polynomial gray tone intensity surface to be more representative of what is happening at the center of the neighborhood than at its periphery. That is, the observed values at the periphery of the neighborhood are likely to deviate more from the corresponding values of the underlying gray tone intensity surface than are the observed values at the center of the neighborhood. Furthermore, we need to pay careful attention to the similarity or dissimilarity of pixels at the periphery of the neighborhood so that their values can be used in a supportive way.

In section 2 we discuss a model and estimation procedure that makes these ideas precise.

2. The Model

Let $\sum_{\substack{ij \\ i+j\leq 3}} \alpha_{ij} r^i c^j$ represent the underlying gray tone intensity surface of a neighborhood, and let $J(r,c)$ represent the observed gray tone values in the neighborhood. At each pixel (r,c) the squared deviation between the representative underlying gray tone intensity surface and the observed image is given by $[\sum_{\substack{ij \\ i+j\leq 3}} \alpha_{ij} r^i c^j - J(r,c)]^2$. The expected value of this squared deviation is the variance of $J(r,c)$ around $\sum_{\substack{ij \\ i+j\leq 3}} \alpha_{ij} r^i c^j$. It is a function of (r,c), and our perspective has suggested that it increases as a monotonic function of the distance between (r,c) and $(0,0)$. We can express this by writing

$$E[\sum_{\substack{ij \\ i+j\leq 3}} \alpha_{ij} r^i c^j - J(r,c)]^2 = \sigma^2[1 + k(r^2+c^2)^P] . \qquad (1)$$

To help make our notation compact, we rewrite this in vector notation. Let J be the vector of observed pixel values in a neighborhood. Let α be the vector of coefficients for the underlying gray tone intensity surface. Let F be a matrix whose columns constitute the discretized polynomial basis. Thus the column corresponding to the basis function $r^i c^j$ has component values that are $r^i c^j$ evaluated at all pixel positions in the neighborhood.

Assuming an ellipsoidally symmetric distribution for the deviations between the observed pixel values and the underlying gray tone intensity surface, we have

$$P(J|F\alpha) = h[(J-F\alpha)' \cancel{\Sigma}_J^{-1}(J-F\alpha)] , \qquad (2)$$

where Σ_J is the covariance matrix of the deviations of the observed values J from the ideal values $F\alpha$.

For the prior distribution of α we likewise take the deviations between the neighborhood α and an α_0 representative of the distribution of α's over all neighborhoods to be distributed in an ellipsoidally symmetric form (typically $\alpha_0 = 0$):

$$P(\alpha) = h[(\alpha-\alpha_0)' \Sigma_\alpha^{-1}(\alpha-\alpha_0)] \; . \qquad (3)$$

From a Bayesian point of view, having observed J we wish to estimate an α that maximizes the probability of α given J. Now,

$$P(\alpha|J) = \frac{P(J|\alpha)P(\alpha)}{P(J)} \; . \qquad (4)$$

Maximizing $P(\alpha|J)$ is equivalent to maximizing $P(J|\alpha)P(\alpha)$, and this is equivalent to maximizing $\log P(J|\alpha) + \log P(\alpha)$. The necessary condition is for the partial derivative of $\log P(J|\alpha) + \log P(\alpha)$ with respect to each component of α to be equal to zero. This yields

$$(-2) \frac{h'[(J-F\alpha)' \Sigma_J^{-1}(J-F\alpha)]}{h[(J-F\alpha)' \Sigma_J^{-1}(J-F\alpha)]} [F' \Sigma_J^{-1}(J-F\alpha)]$$

$$+ (-2) \frac{h'[(\alpha-\alpha_0)' \Sigma_\alpha^{-1}(\alpha-\alpha_0)]}{h[(\alpha-\alpha_0)' \Sigma_\alpha^{-1}(\alpha-\alpha_0)]} \Sigma_\alpha^{-1}(\alpha_0-\alpha) = 0 \; . \qquad (5)$$

In the case where h is the multivariate normal density,

$$h(x^2) = Ae^{-x^2/2} \; . \qquad (6)$$

Or, with a simple argument μ replacing x^2,

$$h(\mu) = Ae^{-\mu/2} \; . \qquad (7)$$

Hence,

$$-2\,\frac{h'(\mu)}{h(\mu)} = -2\,\frac{Ae^{-\mu/2}(-1/2)}{Ae^{-\mu/2}} = 1\,. \tag{8}$$

In the multivariate normal case, the equation simplifies to

$$-F'\,\Sigma_{J}^{-1}(J-F\alpha) + \Sigma_{\alpha}^{-1}(\alpha-\alpha_0) = 0 \tag{9}$$

or

$$[F'\,\Sigma_{J}^{-1}F + \Sigma_{\alpha}^{-1}]\alpha = F'\,\Sigma_{J}^{-1}J + \Sigma_{\alpha}^{-1}\alpha_0\,. \tag{10}$$

To relate this to standard least squares, take $\Sigma_{J}^{-1} = \sigma^2 I$ and $\Sigma_{\alpha}^{-1} = 0$, in which case we have $F'F\alpha = F'J$, which is the usual normal equation. $\Sigma_{\alpha}^{-1} = 0$ means that the variance of α is very large. In essence, it says that nothing is known about α. $\Sigma_{J}^{-1} = \sigma^2 I$ means that the deviations of the observed from the ideal are uncorrelated and that the expected squared deviations are identical throughout the neighborhood rather than increasing for pixels closer to the periphery as suggested earlier.

Now let us move on to a nonnormal case, in which the tails of the distribution are much fatter than the normal distribution. One such distribution is the slash distribution, which arises from a normal (0,1) variate being divided by a uniform (0,1) variate. Another such distribution is the Cauchy distribution.

The slash density function has the form

$$s(x) = \frac{1 - e^{-x^2/2}}{2\pi x^2}\,. \tag{11}$$

Because we have squared the argument before the evaluation, we have

$$s(\mu) = \frac{1 - e^{-\mu/2}}{2\pi\mu}\,, \qquad \mu \geq 0\,. \tag{12}$$

Thus,

$$-\frac{2s'(\mu)}{s(\mu)} = 2\,\frac{1-(1+\mu/2)e^{-\mu/2}}{\mu^2}\,, \tag{13}$$

a function that is always positive, having largest magnitude for small μ and a monotonically decreasing magnitude for larger μ.

The Cauchy distribution has the form

$$c(x) = \frac{1}{\pi(1+x^2)}\,. \tag{14}$$

Because we have squared the argument before evaluation, we have

$$c(x) = \frac{1}{\pi(1+\mu)}\,, \qquad \mu \geq 0\,. \tag{15}$$

Thus,

$$-\frac{c'(\mu)}{c(\mu)} = \frac{1}{1+\mu}\,, \tag{16}$$

a function that is always positive, having largest magnitude for small μ and a monotonically decreasing magnitude for larger μ.

On the basis of the behavior of h'/h for slash and Cauchy distributions, we can discuss the meanings of h'/h in Eq. (5). Simply, if the fit $F\alpha$ to J is relatively good compared to our prior uncertainty about α, then the estimated α is determined mostly by the least squares fit and hardly at all by the prior information we have about α. If the fit $F\alpha$ to J is comparable in uncertainty to our prior uncertainty about α, then the estimated α is determined in equal measure by the least squares fit and by the prior information. If the fit $F\alpha$ to J has more error than our prior uncertainty about α, then the estimated α is determined more by the prior information than by the fit.

To see how this works more precisely, let

$$\lambda_J(\alpha) = \frac{h'[(J-F\alpha)'\Sigma_J^{-1}(J-F\alpha)]}{h[(J-F\alpha)'\Sigma_J^{-1}(J-F\alpha)]} \tag{17}$$

and

$$\lambda_\alpha(\alpha) = \frac{h'[(\alpha-\alpha_0)'\Sigma_\alpha^{-1}(\alpha-\alpha_0)]}{h[(\alpha-\alpha_0)'\Sigma_\alpha^{-1}(\alpha-\alpha_0)]} . \tag{18}$$

Equation (5) becomes

$$[\lambda_J(\alpha)F'\Sigma_J^{-1}F + \lambda_\alpha(\alpha)\Sigma_\alpha^{-1}]\alpha$$

$$= \lambda_J(\alpha)F'\Sigma_J^{-1}J + \lambda_\alpha(\alpha)\Sigma_\alpha^{-1}\alpha_0 . \tag{19}$$

We can solve this equation iteratively. Let α^n be the value of the estimated α at the nth iteration. Take the initial $\alpha^{(1)}$ to satisfy Eq. (10). Suppose $\alpha^{(n)}$ has been determined. Substitute $\alpha^{(n)}$ into Eqs. (17) and (18) to obtain $\lambda_J(\alpha^{(n)})$ and $\lambda_\alpha(\alpha^{(n)})$. Then substitute these values for $\lambda_J(\alpha^n)$ and $\lambda_\alpha(\alpha^n)$ into Eq. (19) to determine $\alpha^{(n+1)}$.

3. The Independence Assumption

An alternative model for the distributions would be for the deviations of the observed values from the values of the underlying gray tone intensity surface to be assumed independent. In this case,

$$P(J/\alpha) = \prod_{(r,c)} P_{rc}(J(r,c)|\alpha)$$

$$= \prod_{(r,c)}^{N} h\left[\left(\frac{J(r,c)^2 - \sum_{n=1} \alpha_n f_n(r,c)}{\sigma_J(r,c)}\right)^2\right] \tag{20}$$

and

$$P(\alpha) = \prod_{n=1}^{N} P_n(\alpha_n | \alpha_{n0})$$

$$= \prod_{n=1}^{N} h\left[\left(\frac{\alpha_n - \alpha_{n0}}{\sigma_{an}}\right)^2\right], \tag{21}$$

where $\alpha' = (\alpha_1,\ldots,\alpha_N)$ and $\alpha_0' = (\alpha_{10},\ldots,\alpha_{N0})$.

Proceeding as before, we obtain that the maximizing α must satisfy

$$\left[F' \Lambda_J \Sigma_J^{-1} F + \Lambda_\alpha \Sigma_\alpha^{-1}\right]\alpha = \left[F' \Lambda_J \Sigma_J^{-1} J + \Lambda_\alpha \Sigma_\alpha^{-1}\right]\alpha_0 , \tag{22}$$

where Σ_J, Σ_α, Λ_J, and Λ_α are diagonal matrices

$$\Sigma_J = \begin{pmatrix} \ddots & & 0 \\ & \sigma_J(r,c)^2 & \\ 0 & & \ddots \end{pmatrix} \tag{23}$$

$$\Sigma_\alpha = \begin{pmatrix} \ddots & & 0 \\ & \sigma_\alpha(n)^2 & \\ 0 & & \ddots \end{pmatrix} \tag{24}$$

$$\Lambda_J = \begin{pmatrix} \ddots & & 0 \\ & \lambda_J(r,c) & \\ 0 & & \ddots \end{pmatrix} \tag{25}$$

$$\Lambda_\alpha = \begin{pmatrix} \ddots & & 0 \\ & \lambda_\alpha(n) & \\ 0 & & \ddots \end{pmatrix} \tag{26}$$

and the diagonal entries of Λ_J and Λ_α are given by

$$\lambda_J(r,c) = \frac{h'\left[\left(\dfrac{J(r,c) - \sum_{n=1}^{N} \alpha_n f_n(r,c)}{\sigma_J(r,c)}\right)^2\right]}{h\left[\left(\dfrac{J(r,c) - \sum_{n=1}^{N} \alpha_n f_n(r,c)}{\sigma_J(r,c)}\right)^2\right]} \tag{27}$$

$$\lambda_\alpha(n) = \frac{h'\left[\left(\dfrac{\alpha_n - \alpha_{n0}}{\sigma_{\alpha n}}\right)^2\right]}{h\left[\left(\dfrac{\alpha_n - \alpha_{n0}}{\sigma_{\alpha n}}\right)^2\right]} . \tag{28}$$

The solution for α can be obtained iteratively. Take the first Λ_J and Λ_α to be the corresponding identity matrices. Solve Eq. (22) for α. Then substitute into Eqs. (27) and (28) for the next Λ_J and Λ_α.

Because the solution for α is iterative, it is not necessary to take the required $NK(1+N+K) + 2N + N^3$ operations to solve Eq. (22) exactly. (The vector J is $K\times 1$ and the vector α is $N\times 1$.) There is a quicker computation procedure. Suppose the basis that is the columns of F satisfies

$$F' \Sigma_J^{-1} F = I . \tag{29}$$

This means that the basis vectors are discretely orthonormal with respect to the weights that are the diagonal entries of the diagonal matrix Σ_J^{-1}. In this case, Eq. (22) holds if and only if

$$(F' \Lambda_J \Sigma_J^{-1} F + \Lambda_\alpha \Sigma_\alpha^{-1} + I - F' \Sigma_J^{-1} F)\alpha$$

$$= F' \Lambda_J \Sigma_J^{-1} J + \Lambda_\alpha \Sigma_\alpha^{-1} \alpha . \tag{30}$$

Rewriting this equation, we have

$$J(I + \Lambda_\alpha \Sigma_\alpha^{-1})\alpha$$

$$= F'[\Lambda_J \Sigma_J^{-1} J + (I - \Lambda_J) \Sigma_J^{-1} F\alpha + \Sigma_J^{-1} F \Lambda_\alpha \Sigma_\alpha^{-1} \alpha_0] \ . \qquad (31)$$

This equation suggests the following iterative procedure for the determination of α:

Take

$$\alpha^{(1)} = (I + \Lambda_\alpha \Sigma_\alpha^{-1})^{-1} (F' \Sigma_J^{-1} J + \Sigma_\alpha^{-1} \alpha_0) \ . \qquad (32)$$

Suppose $\alpha^{(n)}$ has already been determined. Define

$$\alpha^{(n+1)} = (I + \Lambda_\alpha \Sigma_\alpha^{-1})^{-1}$$

$$\cdot \ F' \Sigma_J^{-1} [\Lambda_J J + (I - \Lambda_J) F\alpha(n) + F \Lambda_\alpha \Sigma_\alpha^{-1} \alpha_0] \ . \qquad (33)$$

Each iteration of Eq. (33) requires 3KN+4K+3N operations, and only two to four iterations are necessary to get a reasonably close answer.

4. Robustness

The model assuming the independence of the deviations between the observed values and the underlying gray tone intensity surface is robust. If there are some pixel positions in which J(r,c) deviates greatly from the corresponding value $\sum_{n=1}^{N} \alpha_n f_n(r,c)$ of the underlying gray tone intensity surface, then since $\lambda(r,c)$ is defined by Eq. (27), that is,

$$\lambda_J(r,c) = \frac{h'\left[\left(\dfrac{J(r,c) - \sum_{n=1}^{N} \alpha_n f_n(r,c)}{\sigma_J(r,c)}\right)^2\right]}{h\left[\left(\dfrac{J(r,c) - \sum_{n=1}^{N} \alpha_n f_n(r,c)}{\sigma_J(r,c)}\right)^2\right]} \tag{27}$$

and $-h'/h$ is small for large arguments, $\lambda_J(r,c)$ will be small. To understand the effect of a small $\lambda_J(r,c)$, examine Eq. (33). On the right-hand side of that equation is the expression $\Lambda_J J + (I - \Lambda_J)F\alpha$, which consists of a generalized convex combination of J, a term depending on the observed data, and $F\alpha$, a term depending on the fit to the data. In those components where $\lambda_J(r,c)$ is small, the generalized convex combination tends to ignore $J(r,c)$ and, in effect, to substitute for it the fit $\sum_{n=1}^{N} \alpha_n f_n(r,c)$. Thus small values of $\lambda_J(r,c)$ substitute the fitted values for the observed values. Values of the weight $\lambda_J(r,c)$ close to 1 tend to make the procedure ignore the fitted values and use only the observed values.

The technique is inherently robust. Any observed value that deviates greatly from the fitted value is in a sense ignored and replaced with a fitted value interpolated on the basis of the other pixel values.

To do the λ_J computation, a density function h is required. As we have seen, a normal distributional assumption leads to each λ_J being the same identical constant. Distributional assumptions such as slash or Cauchy lead to λ_J being some monotonically decreasing function of the squared difference between the observed and fitted values. The monotonically decreasing function depends on the distributional assumption being made.

One way to avoid the distributional assumption is to use a form λ_J that has proved to work well over several different kinds of distributions. One such form is Tukey's bisquare function, used in computing the biweight:

$$\lambda_J(r,c) = \begin{cases} [1 - s^2(r,c)]^2, & \text{if } s^2(r,c) < 1 \\ 0, & \text{otherwise} \end{cases} \tag{34}$$

where

$$s^2(r,c) = \frac{[J(r,c) - \sum_{n=1}^{N} \alpha_n f_n(r,c)]^2}{C\,\sigma_J^2(r,c)} \tag{35}$$

and C is a constant with value between 6 and 9. In this case, the estimated coefficients $\alpha_1,\ldots,\alpha_N$ are generalizations of the biweight, and the computational procedure discussed in section 2.1 corresponds to Tukey's iterative reweighted least squares regression procedure [Mosteller and Tukey, 1977].

5. References

Haralick, R. M. (1980), "Edge and region analysis for digital image data," Comput. Graphics and Image Processing 12, pp. 113-129.

Haralick, R. M. (1982), "Zero-crossing of second directional derivative edge operator," Proceedings of the SPIE Technical Symposium East, Arlington, Va., May 3-7, 1982, 336, p. 23.

Haralick, R. M. (1983), "Ridges and valleys on digital images," Comput. Vision Graphics and Image Processing 22, pp. 28-38.

Haralick, R. M. (1984), "Digital step edges from zero-crossing of second directional derivative," IEEE Trans. Pattern Analysis and Machine Intelligence PAMI-6, No. 1, pp. 58-68.

Haralick, R. M., and Layne Watson (1981), "A facet model for image data," Comput. Graphics and Image Processing 15, pp. 113-129.

Mosteller, Frederick, and John Tukey (1977), Data Analysis and Regression, Addison-Wesley, Reading, Mass., pp. 356-358.

Prewitt, Judy (1970), "Object enhancement and extraction," in Picture Processing and Psychopictorics, B. Lipkin and A. Rosenfeld, eds., Academic Press, New York, pp. 75-149.