The Variable Bandwidth Mean Shift and Data-Driven Scale Selection

Dorin Comaniciu Visvanathan Ramesh
Imaging & Visualization Department
Siemens Corporate Research
755 College Road East, Princeton, NJ 08540

Peter Meer
Electrical & Computer Engineering Department
Rutgers University
94 Brett Road, Piscataway, NJ 08855

Abstract

We present two solutions for the scale selection problem in computer vision. The first one is completely nonparametric and is based on the the adaptive estimation of the normalized density gradient. Employing the sample point estimator, we define the Variable Bandwidth Mean Shift, prove its convergence, and show its superiority over the fixed bandwidth procedure. The second technique has a semiparametric nature and imposes a local structure on the data to extract reliable scale information. The local scale of the underlying density is taken as the bandwidth which maximizes the magnitude of the normalized mean shift vector. Both estimators provide practical tools for autonomous image and quasi real-time video analysis and several examples are shown to illustrate their effectiveness.

1 Motivation for Variable Bandwidth

The efficacy of Mean Shift analysis has been demonstrated in computer vision problems such as tracking and segmentation in [5, 6]. However, one of the limitations of the mean shift procedure as defined in these papers is that it involves the specification of a scale parameter. While results obtained appear satisfactory, when the local characteristics of the feature space differs significantly across data, it is difficult to find an optimal global bandwidth for the mean shift procedure. In this paper we address the issue of locally adapting the bandwidth. We also study an alternative approach for data-driven scale selection which imposes a local structure on the data. The proposed solutions are tested in the framework of quasi real-time video analysis.

We review first the intrinsic limitations of the fixed bandwidth density estimation methods. Then, two of the most popular variable bandwidth estimators, the *balloon* and the *sample point*, are introduced and their advantages discussed. We conclude the section by showing that, with some precautions, the performance of the sample point estimator is superior to both fixed bandwidth and balloon estimators.

1.1 Fixed Bandwidth Density Estimation

The multivariate fixed bandwidth kernel density estimate is defined by

$$\hat{f}(\mathbf{x}) = \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{\mathbf{x} - \mathbf{x}_i}{h}\right). \quad (1)$$

where the d-dimensional vectors $\{\mathbf{x}_i\}_{i=1\ldots n}$ represent a random sample from some unknown density f and the kernel, K, is taken to be a radially symmetric, nonnegative function centered at zero and integrating to one. The terminology *fixed bandwidth* is due to the fact that h is held constant across $\mathbf{x} \in R^d$. As a result, the fixed bandwidth procedure (1) estimates the density at each point $\mathbf{x}$ by taking the average of identically scaled kernels centered at each of the data points.

For pointwise estimation, the classical measure of the closeness of the estimator $\hat{f}$ to its target value f is the mean squared error (MSE), equal to the sum of the variance and squared bias

$$\begin{aligned} \mathrm{MSE}(\mathbf{x}) &= E\left[\hat{f}(\mathbf{x}) - f(\mathbf{x})\right]^2 \\ &= \mathrm{Var}\left(\hat{f}(\mathbf{x})\right) + \left[\mathrm{Bias}\left(\hat{f}(\mathbf{x})\right)\right]^2 . \end{aligned} \quad (2)$$

Using the multivariate form of the Taylor theorem, the bias and the variance are approximated by [20, p.97]

$$\mathrm{Bias}(\mathbf{x}) \approx \frac{1}{2} h^2 \mu_2(K) \Delta f(\mathbf{x}) \quad (3)$$

and

$$\mathrm{Var}(\mathbf{x}) \approx n^{-1} h^{-d} R(K) f(\mathbf{x}) , \quad (4)$$

where $\mu_2(K) = \int z_1^2 K(\mathbf{z}) d\mathbf{z}$ and $R(K) = \int K(\mathbf{z}) d\mathbf{z}$ are kernel dependent constants, z_1 is the first component of the vector $\mathbf{z}$, and Δ is the Laplace operator.

The tradeoff of bias versus variance can be observed in (3) and (4). The bias is proportional to h^2, which means that smaller bandwidths give a less biased estimator. However, decreasing h implies an increase in the variance which is proportional to $n^{-1}h^{-d}$. Thus for a fixed bandwidth estimator we should choose h that achieves an optimal compromise between the bias and variance over all $\mathbf{x} \in R^d$, i.e., minimizes the mean integrated squared error (MISE)

$$\mathrm{MISE}(\mathbf{x}) = E \int \left(\hat{f}(\mathbf{x}) - f(\mathbf{x})\right)^2 d\mathbf{x} . \quad (5)$$

Nevertheless, the resulting bandwidth formula (see [17, p.85], [20, p.98]) is of little practical use, since it depends on the Laplacian of the unknown density being estimated.

The best of the currently available data-driven methods for bandwidth selection seems to be the plug-in rule [15], which was proven to be superior to least squares cross validation and biased cross-validation [11], [16,

0-7695-1143-0/01 $10.00 © 2001 IEEE

p.46]. A practical one dimensional algorithm based on this method is described in the Appendix. For the multivariate case, see [20, p.108].

Note that these data-driven bandwidth selectors work well for multimodal data, their only assumption being a certain smoothness in the underlying density. However, the fixed bandwidth affects the estimation performance, by undersmoothing the tails and oversmoothing the peaks of the density. The performance also decreases when the data exhibits local scale variations.

1.2 Balloon and Sample Point Estimators

According to expression (1), the bandwidth h can be varied in two ways. First, by selecting a different bandwidth $h = h(\mathbf{x})$ for each *estimation point* $\mathbf{x}$, one can define the *balloon* density estimator

$$\hat{f}_1(\mathbf{x}) = \frac{1}{nh(\mathbf{x})^d} \sum_{i=1}^{n} K\left(\frac{\mathbf{x}-\mathbf{x}_i}{h(\mathbf{x})}\right). \tag{6}$$

In this case, the estimate of f at $\mathbf{x}$ is the average of identically scaled kernels centered at each data point.

Second, by selecting a different bandwidth $h = h(\mathbf{x}_i)$ for each *data point* $\mathbf{x}_i$ we obtain the *sample point* density estimator

$$\hat{f}_2(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{m} \frac{1}{h(\mathbf{x}_i)^d} K\left(\frac{\mathbf{x}-\mathbf{x}_i}{h(\mathbf{x}_i)}\right). \tag{7}$$

for which the estimate of f at $\mathbf{x}$ is the average of differently scaled kernels centered at each data point.

While the balloon estimator has more intuitive appeal, its performance improvement over the fixed bandwidth estimator is insignificant. When the bandwidth $h(\mathbf{x})$ is chosen as a function of the k-th nearest neighbor, the bias and variance are still proportional to h^2 and $n^{-1}h^{-d}$, respectively [8]. In addition, the balloon estimators usually fail to integrate to one.

The sample point estimators, on the other hand, are themselves densities, being non-negative and integrating to one. Their most attractive property is that a particular choice of $h(\mathbf{x}_i)$ reduces considerably the bias. Indeed, when $h(\mathbf{x}_i)$ is taken to be reciprocal to the square root of $f(\mathbf{x}_i)$

$$h(\mathbf{x}_i) = h_0 \left[\frac{\lambda}{f(\mathbf{x}_i)}\right]^{1/2} \tag{8}$$

the bias becomes proportional to h^4, while the variance remains unchanged, proportional to $n^{-1}h^{-d}$ [1, 8]. In (8), h_0 represents a fixed bandwidth and λ is a proportionality constant.

Since $f(\mathbf{x}_i)$ is unknown it has to be estimated from the data. The practical approach is to use one of the methods described in Section 1.1 to find h_0 and an initial estimate (called *pilot*) of f denoted by $\tilde{f}$. Note that by using $\tilde{f}$ instead of f in (8), the nice properties of the sample point estimators (7) remain unchanged [8]. Various authors [16, p.56], [17, p.101] remarked that the method is insensitive to the fine detail of the pilot estimate. The only provision that should be taken is to bound the pilot density away from zero.

The final estimate (7) is however influenced by the choice of the proportionality constant λ, which divides the range of density values into *low* and *high* densities. When the local density is low, i.e., $\tilde{f}(\mathbf{x}_i) < \lambda$, $h(\mathbf{x}_i)$ increases relative to h_0 implying more smoothing for the point $\mathbf{x}_i$. For data points that verify $\tilde{f}(\mathbf{x}_i) > \lambda$, the bandwidth becomes narrower.

A good initial choice [17, p.101] is to take λ as the geometric mean of $\left\{\tilde{f}(x_i)\right\}_{i=1\ldots n}$. Our experiments have shown that for superior results, a certain degree of tuning is required for λ. Nevertheless, the sample point estimator proved to be almost all the time much better than the fixed bandwidth estimator.

2 Variable Bandwidth Mean Shift

We show next that starting from the sample point estimator (7) an adaptive estimator of the density's normalized gradient can be defined. The new estimator, which associates to each data point a differently scaled kernel, is the basic step for an iterative procedure that we prove to converge to a local mode of the underlying density, when the kernel obeys some mild constraints. We called the new procedure the *Variable Bandwidth Mean Shift*. Due to its excellent statistical properties, we anticipate the extensive use of the adaptive estimator by vision applications that require minimal human intervention.

2.1 Definitions

To simplify notations we proceed as in [6] by introducing first the *profile* of a kernel K as a function $k : [0, \infty) \rightarrow R$ such that $K(\mathbf{x}) = k(\|\mathbf{x}\|^2)$. We also denote $h_i \equiv h(\mathbf{x}_i)$ for all $i = 1 \ldots n$. Then, the sample point estimator (7) can be written as

$$\hat{f}_K(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h_i^d} k\left(\left\|\frac{\mathbf{x}-\mathbf{x}_i}{h_i}\right\|^2\right), \tag{9}$$

where the subscript K indicates that the estimator is based on kernel K.

A natural estimator of the gradient of f is the gradient of $\hat{f}_K(\mathbf{x})$

$$\begin{aligned}
\hat{\nabla} f_K(\mathbf{x}) \equiv \nabla \hat{f}_K(\mathbf{x}) &= \frac{2}{n} \sum_{i=1}^{n} \frac{\mathbf{x}-\mathbf{x}_i}{h_i^{d+2}} k'\left(\left\|\frac{\mathbf{x}-\mathbf{x}_i}{h_i}\right\|^2\right) \\
&= \frac{2}{n} \sum_{i=1}^{n} \frac{\mathbf{x}_i-\mathbf{x}}{h_i^{d+2}} g\left(\left\|\frac{\mathbf{x}-\mathbf{x}_i}{h_i}\right\|^2\right) \\
&= \frac{2}{n}\left[\sum_{i=1}^{n} \frac{1}{h_i^{d+2}} g\left(\left\|\frac{\mathbf{x}-\mathbf{x}_i}{h_i}\right\|^2\right)\right] \times
\end{aligned}$$

$$\times \left[\frac{\sum_{i=1}^{n} \frac{\mathbf{x}_i}{h_i^{d+2}} g\left(\left\| \frac{\mathbf{x}-\mathbf{x}_i}{h_i} \right\|^2 \right)}{\sum_{i=1}^{n} \frac{1}{h_i^{d+2}} g\left(\left\| \frac{\mathbf{x}-\mathbf{x}_i}{h_i} \right\|^2 \right)} - \mathbf{x} \right], \quad (10)$$

where we denoted

$$g(x) = -k'(x) , \quad (11)$$

and assumed that the derivative of profile k exists for all $x \in [0, \infty)$, except for a finite set of points.

The last bracket in (10) represents the variable bandwidth mean shift vector

$$M_v(\mathbf{x}) \equiv \frac{\sum_{i=1}^{n} \frac{\mathbf{x}_i}{h_i^{d+2}} g\left(\left\| \frac{\mathbf{x}-\mathbf{x}_i}{h_i} \right\|^2 \right)}{\sum_{i=1}^{n} \frac{1}{h_i^{d+2}} g\left(\left\| \frac{\mathbf{x}-\mathbf{x}_i}{h_i} \right\|^2 \right)} - \mathbf{x} \quad (12)$$

To see the significance of expression (12), we define first the kernel G as

$$G(\mathbf{x}) = Cg(\|\mathbf{x}\|^2), \quad (13)$$

where C is a normalization constant that forces G to integrate to one.

Then, by employing (8), the term that multiplies the mean shift vector in (10) can be written as

$$\frac{2}{n}\left[\sum_{i=1}^{n} \frac{1}{h_i^{d+2}} g\left(\left\| \frac{\mathbf{x}-\mathbf{x}_i}{h_i} \right\|^2 \right) \right] = \frac{2}{C}\left[\frac{\sum_{i=1}^{n} \tilde{f}(\mathbf{x}_i)}{n\lambda h_0^2} \right] \hat{f}_G(\mathbf{x}) \quad (14)$$

where

$$\hat{f}_G(\mathbf{x}) \equiv C \frac{\sum_{i=1}^{n} \tilde{f}(\mathbf{x}_i) \frac{1}{h_i^d} g\left(\left\| \frac{\mathbf{x}-\mathbf{x}_i}{h_i} \right\|^2 \right)}{\sum_{i=1}^{n} \tilde{f}(\mathbf{x}_i)} \quad (15)$$

is nonnegative and integrates to one, representing an estimate of the density of the data points weighted by the pilot density values $\tilde{f}(\mathbf{x}_i)$.

Finally, by using (10), (12), and (14) it results that

$$M_v(\mathbf{x}) = \frac{\lambda}{n^{-1}\sum_{i=1}^{n} \tilde{f}(\mathbf{x}_i)} \frac{h_0^2}{2/C} \frac{\hat{\nabla} f_K(\mathbf{x})}{\hat{f}_G(\mathbf{x})}. \quad (16)$$

Equation (16) represents a generalization of equation (13) derived in [6] for the fixed bandwidth mean shift. It shows that the adaptive bandwidth mean shift is an estimator of the normalized gradient of the underlying density.

The proportionality constant, however, depends on the value of λ. When λ is increased, the norm of the mean shift vector also increases. On the other hand, a small value for λ implies a small $\|M_v\|$. Due to this external variability of the mean shift norm, the convergence property of an iterative procedure based on the variable bandwidth mean shift is remarkable. Note also that when λ is taken equal to the arithmetic mean of $\left\{\tilde{f}(x_i)\right\}_{i=1...n}$, the proportionality constant becomes as in the fixed bandwidth case.

2.2 Properties of the Adaptive Mean Shift

Equation (12) shows an attractive behavior of the adaptive estimator. The data points lying in large density regions affect a narrower neighborhood since the kernel bandwidth h_i is smaller, but are given a larger importance, due to the weight $1/h_i^{d+2}$. By contrast, the points that correspond to the tails of the underlying density are smoothed more and receive a smaller weight. The extreme points (outliers) receive very small weights, being thus automatically discarded. Recall that the fixed bandwidth mean shift [5, 6] associates the same kernel for each data point.

The most important property of the adaptive estimator is the convergence associated with its repetitive computation. In other words, if we define the *mean shift procedure* recursively as the evaluation of the mean shift vector $M_v(\mathbf{x})$ followed by the translation of the kernel G by $M_v(\mathbf{x})$, this procedure leads to a stationary point (zero gradient) of the underlying density. More specifically, we will show that the point of convergence represents a stationary point of the sample point estimator (9). Thus, the superior performance of the sample point estimator translates into superior performance for the adaptive mean shift.

We denote by $\{\mathbf{y}_j\}_{j=1,2...}$ the sequence of successive locations of the kernel G, where

$$\mathbf{y}_{j+1} = \frac{\sum_{i=1}^{n} \frac{\mathbf{x}_i}{h_i^{d+2}} g\left(\left\| \frac{\mathbf{y}_j-\mathbf{x}_i}{h_i} \right\|^2 \right)}{\sum_{i=1}^{n} \frac{1}{h_i^{d+2}} g\left(\left\| \frac{\mathbf{y}_j-\mathbf{x}_i}{h_i} \right\|^2 \right)}, \quad j = 1, 2, \ldots \quad (17)$$

is the weighted mean at $\mathbf{y}_j$ computed with kernel G and weights $1/h_i^{d+2}$, and $\mathbf{y}_1$ is the center of the initial kernel. The density estimates computed with kernel K in the points (17) are

$$\hat{f}_K = \left\{ \hat{f}_K(j) \right\}_{j=1,2...} \equiv \left\{ \hat{f}_K(\mathbf{y}_j) \right\}_{j=1,2...} . \quad (18)$$

We show in Appendix that if the kernel K has a convex and monotonic decreasing profile and the kernel G is defined according to (11) and (13), the sequences (17) and (18) are convergent. This means that the adaptive mean shift procedure initialized at a given location, converges at a nearby point where the estimator (9) has zero gradient. In addition, since the modes of the density are points of zero gradient, it results that the convergence point is a mode candidate.

The advantage of using the mean shift rather than the direct computation of (9) followed by a search for local maxima is twofold. First, the overall computational complexity of the mean shift is much smaller than that of the direct method. The direct search for maxima requires a number of density function evaluations that increases exponentially with the space dimension. Second, for many applications (see for example [6]) we only

need to know the mode associated with a reduced set of data points. In this case, the mean shift procedure becomes a natural process that follows the trail to the local mode.

The iterative procedure for mode detection based on the variable bandwidth mean shift is summarized below.

Variable Bandwidth Mean Shift Algorithm

Given the data points $\{\mathbf{x}_i\}_{i=1\ldots n}$:

1. Derive a fixed bandwidth h_0 and a pilot estimate $\tilde{f}$ using the plug-in rule (see Appendix for the one dimensional plug-in rule).
2. Compute $\log\lambda = n^{-1}\sum_{i=1}^{n}\log\tilde{f}(\mathbf{x}_i)$.
3. For each data point $\mathbf{x}_i$ compute its adaptive bandwidth $h(\mathbf{x}_i) = h_0\left[\lambda/\tilde{f}(\mathbf{x}_i)\right]^{1/2}$.
4. Initialize $\mathbf{y}_1$ with the location of interest and compute iteratively (17) till convergence. The convergence point is a point of zero gradient, hence, a mode candidate.

2.3 Performance Comparison

We compared the variable and fixed bandwidth mean shift algorithms for various multimodal data sets that exhibited also scale variations. The fixed bandwidth procedure was run with a bandwidth h_0 derived from the plug-in rule given in Appendix.

The plug-in rule was developed for density estimation [15] and since here we are concerned with density gradient estimation it is recommended [20, p.49] to use a larger bandwidth to compensate for the inherently increased sensitivity of the estimation process. We have modified the plug-in rule by halvening the contribution of the variance term. This change was maintained for all the experiments presented in this paper. The constant λ of the adaptive procedure was kept as the geometric mean of $\left\{\tilde{f}(x_i)\right\}_{i=1\ldots n}$.

As one can see from Figures 1 and 2 the fixed bandwidth mean shift resulted in good performance for the locations where the local scale was in the medium range. However, the very narrow peaks were fused, while the tails were broken into pieces. On the other hand, the adaptive algorithm showed superior performance, by choosing a proper bandwidth for each data point.

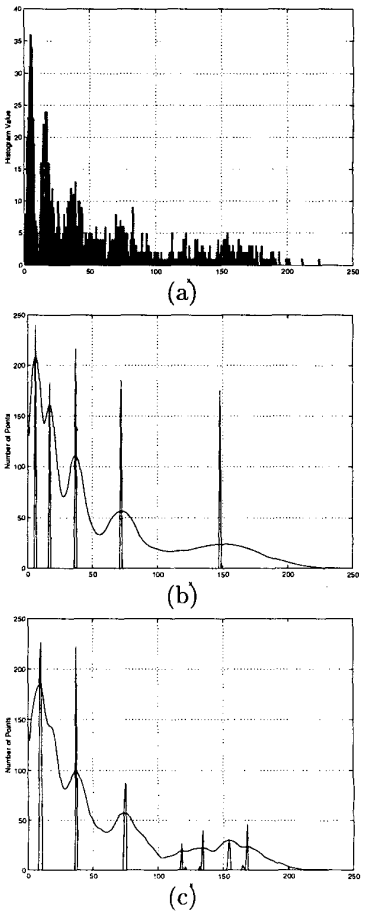


Figure 1: A mixture of 200 data points from each N(5,2), N(17,4), N(37,8), N(70,16), N(145,32). The continuous line is a scaled version of the density estimate. The detected modes are marked proportional to the number of data points that converged to them. (a) Histogram of the data. (b) Variable Bandwidth. (c) Fixed Bandwidth.

3 Semiparametric Scale Selection

3.1 Motivation

The previous two sections followed purely nonparametric ideas, since no formal structure was assumed about the data. Implying only a certain smoothness of the underlying density we used available algorithms for scale selection to derive an initial bandwidth h_0. The criterion for bandwidth selection was a global measure (MISE), hence, h_0 achieved an optimal compromise between the integrated squared bias and the integrated variance. Then, we modified this bandwidth for each data point, according to the local density.

The main problem with this approach is that for multidimensional multimodal data, it is very difficult to determine the right h_0 from the sample points and many of the practical issues are yet to be resolved [20, p.108]. As a consequence, most of practical algorithms use empirical bandwidth selection rules that are less dependent or even independent from the sample data. This implies a decrease in their performance when the input statistics is nonstationary, as it happens most of

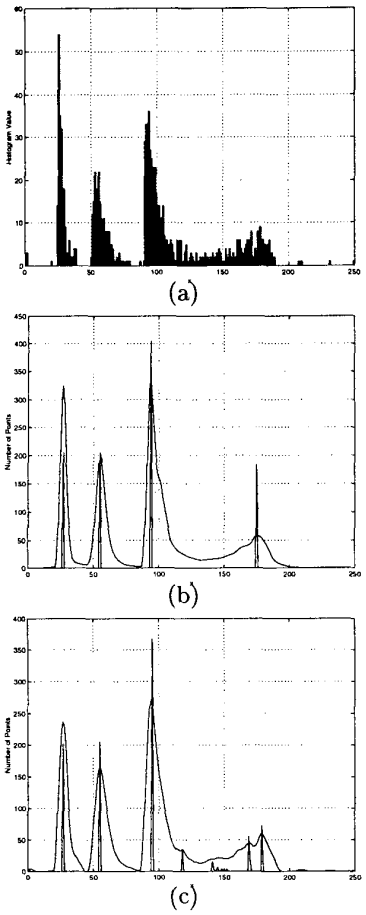


Figure 2: A mixture of 200 data points from each exp(3)+25, 2chi2(4)+50, lognormal(2,1)+90, lognormal(2,1)+90, 190-lognormal(3,1). The continuous line is a scaled version of the density estimate. The detected modes are marked proportional to the number of data points that converged to them. (a) Histogram of the data. (b) Variable Bandwidth. (c) Fixed Bandwidth.

the time in vision tasks.

3.2 Normalized Mean Shift Based Scale Selection

We propose in this section a different approach for bandwidth selection. The idea is to impose a local structure on the data by assuming that *locally* the underlying density is spherical normal with unknown mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{I}$.

At a first look, the task of finding $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ for each data point seems to be very difficult. To locally fit a normal to the multivariate data one needs a priori knowledge of the neighborhood size in which the unknown parameters are to be estimated. If the estimation is performed for several neighborhood sizes, a scale invariant measure of the goodness of fit is needed.

Fortunately, a simple solution exists. It is based on the following theorem, valid when the number of available samples is large.

Theorem 1 *If the true density f is normal with parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{I}$, and the fixed bandwidth mean shift is computed with a spherical normal kernel of bandwidth h_0, then, the bandwidth normalized norm of the mean shift vector is maximized when $h_0 \equiv \sigma$.*

Proof Recall that the fixed bandwidth mean shift vector computed with kernel G of bandwidth h_0 can be written as

$$M(\mathbf{x}) = \frac{h_0^2}{2/C} \frac{\nabla \hat{f}_K(\mathbf{x})}{\hat{f}_G(\mathbf{x})}. \tag{19}$$

Since the true density f is normal with covariance matrix $\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{I}$ it follows that the mean of $\hat{f}_G(\mathbf{x})$, $E\left[\hat{f}_G(\mathbf{x})\right] \equiv \phi(\mathbf{x}; \sigma^2 + h_0^2)$ is also a normal surface with covariance $(\sigma^2 + h_0^2)\boldsymbol{I}$. Likewise, by taking into account (11) we have $E\left[\nabla \hat{f}_K(\mathbf{x})\right] = \nabla \phi(\mathbf{x}; \sigma^2 + h_0^2)$.

By assuming that the large sample approximation is valid (see [18]) it results that

$$\begin{aligned} \text{plim} M(\mathbf{x}) &= \frac{h_0^2}{2/C} \frac{E\left[\nabla \hat{f}_K(\mathbf{x})\right]}{E\left[\hat{f}_G(\mathbf{x})\right]} = \frac{h_0^2}{2/C} \frac{\nabla \phi(\mathbf{x}; \sigma^2 + h_0^2)}{\phi(\mathbf{x}; \sigma^2 + h_0^2)} \\ &= -\frac{1}{2/C} \frac{h_0^2}{\sigma^2 + h_0^2} (\mathbf{x} - \boldsymbol{\mu}), \end{aligned} \tag{20}$$

where plim denotes probability limit with h_0 held constant. This is equivalent to assuming the sample size sufficiently large to make the variances of the means relatively small.

Finally, the norm of the bandwidth normalized mean shift is

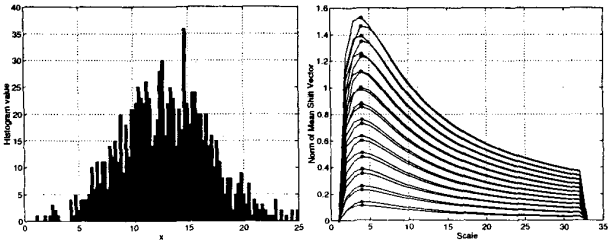
$$\left\| \frac{\text{plim} M(\mathbf{x})}{h_0} \right\| = \frac{1}{2/C} \frac{h_0}{\sigma^2 + h_0^2} \|\mathbf{x} - \boldsymbol{\mu}\|, \tag{21}$$

a quantity that has a unique positive maximum at $h_0 = \sigma$.

Theorem 1 leads to a very simple and accurate scale selection rule: the underlying density has the local scale equal to the bandwidth that maximizes the norm of the normalized mean shift vector. We expect that a similar property holds in the case of anisotropic covariance matrices.

3.3 Scale Selection Experiments

Figure 3a shows a data set of size $n = 2000$, drawn from N(4,10). The bandwidth normalized mean shift is represented in Figure 3b as a function of scale. Observe



(a) (b)

Figure 3: Semiparametric scale selection. (a) Input data. N(10,4), $n = 2000$. (b) Normalized mean shift as a function of scale for the points with positive mean shift. The upper curves correspond to the points located far from the mean. The curves are maximized for $h_0 = 4$.

the accurate local scale indication by the maxima of the curves. The same accurate results were obtained for two and three dimensions.

4 Video Data Analysis

A fundamental task in video data analysis is to detect *blobs* represented by collections of pixels that are coherent in spatial, range, and time domain [21]. The two dimensional space of the lattice is known as the *spatial* domain while the gray level, color, spectral, or texture information is represented in the *range* domain.

Based on the two new estimators introduced in Sections 2 and 3 we present next an autonomous technique that segment a video frame into representative blobs detected in the spatial and color domains. The technique can be naturally extended to incorporate time information, this being one of the subjects of our current work.

We selected the orthogonal features $I1 = (R+G+B)/3$, $I2 = (R-B)/2$ and $I3 = (2G-R-B)/4$ from [10] to represent the color information. Due to the orthogonality of the features, the one dimensional plug-in rule for bandwidth selection can be applied independently for each color coordinate.

As in [5], the idea is to apply the mean shift procedure for the data points in the joint spatial-range domain. Each data point becomes associated to a point of convergence which represents the local mode of the density in a $d = 2 + 3$ dimensional space (2 spatial components and 3 color components).

We employed a spherical kernel for the spatial domain and a product kernel for the three color components. The efficiency of the product kernel is known to be very close to that of spherical kernels [20, p.104].

Due to the different nature of the two spaces, the problem of bandwidth selection has been treated differently for each space. A fixed bandwidth was first derived for each color component, based on the one dimensional plug-in rule. Then, the pilot density has been computed for each pixel, and the adaptive color bandwidths were determined according to (8) for each pixel. This process has been repeated for different scales of the spatial kernel. Finally, the spatial scale has been selected for each pixel according to the semiparametric rule. As a result, each pixel received a unique color bandwidth for color and a unique spatial bandwidth.

To obtain the segmented image, the adaptive mean shift procedure has been applied in the joint domain. The blobs were identified as groups of pixels that had the same connected convergence points (see [5]). The algorithm is summarized below.

Adaptive Mean Shift Segmentation

Given the image pixels $\{\mathbf{x}_i; I1_i, I2_i, I3_i\}_{i=1\ldots n}$, and a range of spatial scales $r_1 \ldots r_S$:

1. Derive h_1, h_2, h_3, a fixed bandwidth for each color feature.
2. For the spatial scale r_1, compute the adaptive bandwidths $h_1(\mathbf{x}_i; r_1)$, $h_2(\mathbf{x}_i; r_1)$, $h_3(\mathbf{x}_i; r_1)$ and determine the magnitude of the normalized mean shift vector $M(\mathbf{x}_i; r_1)$.
3. Repeat Step 2. for the spatial scales $r_2 \ldots r_S$.
4. Select for each pixel a spatial scale r_j according to the semiparametric rule. Select also the color bandwidths $h_1(\mathbf{x}_i; r_j)$, $h_2(\mathbf{x}_i; r_j)$, and $h_3(\mathbf{x}_i; r_j)$.
5. Run the adaptive mean shift procedure, and identify the blobs as groups of pixels having the same connected convergence points.

Although the adaptive algorithm has an increased complexity, its careful software implementation with three spatial scales (S=3) runs at about 8 frames/second on a Dual Pentium III at 900MHz for a video frame size of 320×240 pixels. Figure 4 shows four examples demonstrating the segmentation of color image data with very different statistics. Figure 5 shows the stability of the algorithm in segmenting a color sequence obtained by panning the camera. The identified blobs were maintained very stable, although the scene data changed gradually along with the camera gain.

5 Discussion

The most attractive property of the techniques proposed in this paper is the automatic bandwidth selection in both color and spatial domain.

The reason we used two different bandwidth selection techniques for the two spaces was not arbitrary.

While the color information can be collected across the image, allowing the computation of robust initial bandwidth for color, the spatial properties of the blobs vary drastically across the image, requiring local decisions for spatial scale selection.

The process defined by the mean shift technique in the color domain resembles bilateral filtering [19] (see also [3] for a discussion on the link between bilateral filtering, anisotropic diffusion [12], and adaptive smoothing [13]). Due to the weighting of the data, the adaptive bandwidth mean shift is more related to robust anisotropic diffusion [4].

In the spatial domain, the mean shift is close to multiscale techniques such as [2], and the semiparametric scale selection rule resembles in principle to those developed in [7, 9].

The unification of all these ideas is an interesting subject for further research.

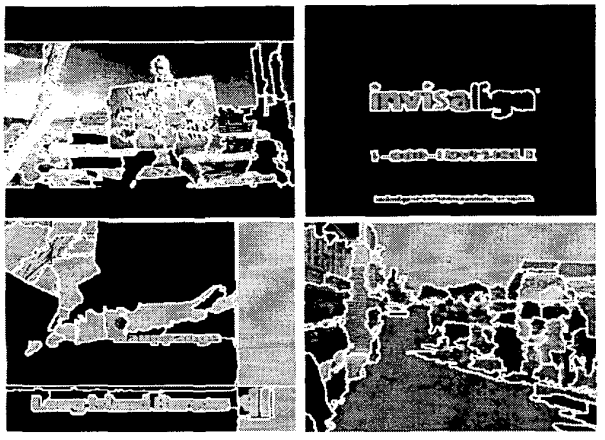


Figure 4: Segmentation examples. Frame size: 320 × 240 pixels.

APPENDIX

One dimensional plug-in rule [15]

1. Compute $\hat{\gamma} = Q_3 - Q_1$, the sample interquartile range.
2. Compute $a = 0.920\hat{\gamma}n^{-1/7}$, $b = 0.912\hat{\gamma}n^{-1/9}$.
3. $\hat{T}_D(b) = -\{n(n-1)\}^{-1}b^{-7}\sum_{i=1}^{n}\sum_{j=1}^{n}\phi^{vi}\{b^{-1}(\mathbf{x}_i-\mathbf{x}_j)\}$

   where ϕ^{vi} is the sixth derivative of the normal kernel (see [20] [p.177]).
4. $\hat{S}_D(a) = \{n(n-1)\}^{-1}a^{-5}\sum_{i=1}^{n}\sum_{j=1}^{n}\phi^{iv}\{a^{-1}(\mathbf{x}_i-\mathbf{x}_j)\}$,

   where ϕ^{iv} is the forth derivative of the normal kernel.

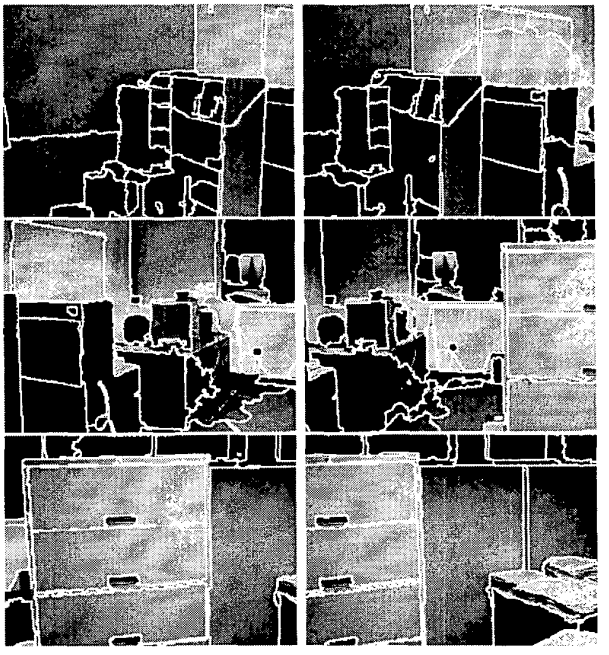


Figure 5: Sequence of segmented images used to test the stability of our algorithm. Frame size: 320 × 240 pixels.

5. $\hat{\alpha}_2(h) = 1.357\{\hat{S}_D(a)/\hat{T}_D(b)\}^{1/7}h^{5/7}$.
6. Solve the equation in h

   $$[R(K)/\{\mu_2^2(K)\hat{S}_D(\hat{\alpha}_2(h))\}]^{1/5}n^{-1/5} - h = 0,$$

   where $\mu_2(K)$ and $R(K)$ are defined in (3) and (4), respectively.

Convergence Proof for Variable Bandwidth Mean Shift

Since n is finite the sequence $\hat{f}_K$ is bounded, therefore, it is sufficient to show that $\hat{f}_K$ is strictly monotonic increasing, i.e., if $\mathbf{y}_j \neq \mathbf{y}_{j+1}$ then $\hat{f}_K(j) < \hat{f}_K(j+1)$, for all $j = 1, 2 \ldots$.

By assuming without loss of generality that $\mathbf{y}_j = \mathbf{0}$ we write

$$\hat{f}_K(j+1) - \hat{f}_K(j) = \\ = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h_i^d}\left[k\left(\left\|\frac{\mathbf{y}_{j+1}-\mathbf{x}_i}{h_i}\right\|^2\right) - k\left(\left\|\frac{\mathbf{x}_i}{h_i}\right\|^2\right)\right] \quad \text{(B.1)}$$

The convexity of the profile k implies that

$$k(x_2) \geq k(x_1) + k'(x_1)(x_2 - x_1) \quad \text{(B.2)}$$

for all $x_1, x_2 \in [0, \infty)$, $x_1 \neq x_2$, and since $k' = -g$, the inequality (B.2) becomes

$$k(x_2) - k(x_1) \geq g(x_1)(x_1 - x_2). \quad \text{(B.3)}$$

Using now (B.1) and (B.3) we have

$$\begin{aligned}
&\hat{f}_K(j+1) - \hat{f}_K(j) \geq \\
&\geq \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h_i^{d+2}} g\left(\left\|\frac{\mathbf{x}_i}{h_i}\right\|^2\right)\left[\|\mathbf{x}_i\|^2 - \|\mathbf{y}_{j+1} - \mathbf{x}_i\|^2\right] \\
&= \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h_i^{d+2}} g\left(\left\|\frac{\mathbf{x}_i}{h_i}\right\|^2\right)\left[2\mathbf{y}_{j+1}^{\top}\mathbf{x}_i - \|\mathbf{y}_{j+1}\|^2\right] \\
&= \frac{1}{n} 2\mathbf{y}_{j+1}^{\top} \sum_{i=1}^{n} \frac{x_i}{h_i^{d+2}} g\left(\left\|\frac{\mathbf{x}_i}{h_i}\right\|^2\right) - \\
&- \frac{1}{n}\|\mathbf{y}_{j+1}\|^2 \sum_{i=1}^{n} \frac{1}{h_i^{d+2}} g\left(\left\|\frac{\mathbf{x}_i}{h_i}\right\|^2\right)
\end{aligned} \tag{B.4}$$

and by employing (17) it results that

$$\hat{f}_K(j+1) - \hat{f}_K(j) \geq \frac{1}{n}\|\mathbf{y}_{j+1}\|^2 \sum_{i=1}^{n} \frac{1}{h^{d+2}} g\left(\left\|\frac{\mathbf{x}_i}{h_i}\right\|^2\right). \tag{B.5}$$

Since k is monotonic decreasing we have $-k'(x) \equiv g(x) \geq 0$ for all $x \in [0, \infty)$. The sum $\sum_{i=1}^{n} \frac{1}{h^{d+2}} g\left(\left\|\frac{\mathbf{x}_i}{h_i}\right\|^2\right)$ is strictly positive, since it was assumed to be nonzero in the definition of the mean shift vector (12). Thus, as long as $\mathbf{y}_{j+1} \neq \mathbf{y}_j = \mathbf{0}$, the right term of (B.5) is strictly positive, i.e., $\hat{f}_K(j+1) - \hat{f}_K(j) > 0$. Hence, the sequence $\hat{f}_K$ is convergent.

To show the convergence of the sequence $\{\mathbf{y}_j\}_{j=1,2...}$ we rewrite (B.5) but without assuming that $\mathbf{y}_j = \mathbf{0}$. After some algebra it results that

$$\hat{f}_K(j+1) - \hat{f}_K(j) \geq \frac{1}{n}\|\mathbf{y}_{j+1} - \mathbf{y}_j\|^2 \sum_{i=1}^{n} \frac{1}{h_i^{d+2}} g\left(\left\|\frac{\mathbf{y}_j - \mathbf{x}_i}{h_i}\right\|^2\right) \tag{B.6}$$

Since $\hat{f}_K(j+1) - \hat{f}_K(j)$ converges to zero, (B.6) implies that $\|\mathbf{y}_{j+1} - \mathbf{y}_j\|$ also converges to zero, i.e., $\{\mathbf{y}_j\}_{j=1,2...}$ is a Cauchy sequence. But any Cauchy sequence is convergent in the Euclidean space, therefore, $\{\mathbf{y}_j\}_{j=1,2...}$ is convergent.

Acknowledgment

Peter Meer was supported by the NSF under the grant IRI 99-87695.

References

[1] I.S. Abramson, "On Bandwidth Variation in Kernel Estimates - A Square Root Law," *The Annals of Statistics*, 10(4):1217–1223, 1982.

[2] N. Ahuja, "A Transform for Multiscale Image Segmentation by Integrated Edge and Region Detection," *IEEE Trans. Pattern Anal. Machine Intell.*, 18:1211–1235, 1996.

[3] D. Barash, "Bilateral Filtering and Anisotropic Diffusion: Towards a Unified Viewpoint," *Hewlett-Packard HPL-2000-18(R.1)*. Available at *http:www.hpl.hp.com.*

[4] M.J. Black, G. Sapiro, D.H. Marimont, D. Heeger,, "Robust Anisotropic Diffusion," *Image Processing*, 7(3):421–432, 1998.

[5] D. Comaniciu, P. Meer, "Mean Shift Analysis and Applications," *IEEE Int'l Conf. Comp. Vis.*, Kerkyra, Greece, 1197–1203, 1999.

[6] D. Comaniciu, V. Ramesh, P. Meer, "Real-Time Tracking of Non-Rigid Objects using Mean Shift," *IEEE Conf. Comp. Vis. Patt. Recogn.*, Hilton Head, South Carolina, Vol. 2, 142–149, 2000.

[7] J. Elder, S.W. Zucker, "Local Scale Control for Edge Detection and Blur Estimation," *IEEE Trans. Pattern Anal. Machine Intell.*, 20(7):699–716, 1998.

[8] P. Hall, T.C. Hui, J.S. Marron, "Improved Variable Window Kernel Estimates of Probability Densities," *The Annals of Statistics*, 23(1):1–10, 1995.

[9] T. Lindeberg, "Edge Detection and Ridge Detection with Automatic Scale Selection," *Int. J. Comp. Vision.*, 30(2):117–154, 1998.

[10] Y. Ohta, T. Kanade, T. Sakai, "Color Information for Region Segmentation," *Computer Graphics and Image Processing*, 13:222–241, 1980.

[11] B. Park, J.S. Marron, "Comparison of Data-Driven Bandwidth Selectors," *J. Am. Statist. Assoc.*, 85(409):66–72, 1990.

[12] P. Perona, J. Malik, "Scale-Space and Edge Detection Using Anisotropic Diffusion," *IEEE Trans. Pattern Anal. Machine Intell.*, 12(7):629–639, 1990.

[13] P. Saint-Marc, J.S. Chen, G.G. Medioni, "Adaptive Smoothing: A General Tool for Early Vision", *IEEE Trans. PAMI*, 13(7):514-529, 1991.

[14] D.W. Scott, *Multivariate Density Estimation*, New York: Wiley, 1992.

[15] S.J. Sheather, M.C. Jones, "A Reliable Data-based Bandwidth Selection Method for Kernel Density Estimation," *J. R. Statist. Soc. B*, 53(3):683–690, 1991.

[16] J.S. Simonoff, *Smoothing Methods in Statistics*, New York: Springer-Verlag, 1996.

[17] B.W. Silverman, *Density Estimation for Statistics and Data Analysis*, New York: Chapman and Hall, 1986.

[18] T.M Stocker, "Smoothing Bias in Density Derivative Estimation," *American Stat. Assoc.*, 88(423):855–863, 1993.

[19] C. Tomasi, R. Manduchi, "Bilateral Filtering for Gray and Color Images", *Int'l Conf. Comp. Vis.*, Bombay, India, 839–846, 1998.

[20] M.P. Wand, M.C. Jones, *Kernel Smoothing*, London: Chapman & Hall, 1995.

[21] C. Wren, A. Azarbayejani, T. Darrell, A. Pentland, "Pfinder: Real-Time Tracking of the Human Body," *IEEE Trans. Pattern Analysis Machine Intell.*, 19:780–785, 1997.