

Mean Shift: A Robust Approach Toward Feature Space Analysis

Dorin Comaniciu, *Member, IEEE*, and Peter Meer, *Senior Member, IEEE*

Abstract—A general nonparametric technique is proposed for the analysis of a complex multimodal feature space and to delineate arbitrarily shaped clusters in it. The basic computational module of the technique is an old pattern recognition procedure, the mean shift. We prove for discrete data the convergence of a recursive mean shift procedure to the nearest stationary point of the underlying density function and, thus, its utility in detecting the modes of the density. The relation of the mean shift procedure to the Nadaraya-Watson estimator from kernel regression and the robust M-estimators of location is also established. Algorithms for two low-level vision tasks, discontinuity preserving smoothing and image segmentation, are described as applications. In these algorithms, the only user set parameter is the resolution of the analysis and either gray level or color images are accepted as input. Extensive experimental results illustrate their excellent performance.

Index Terms—Mean shift, clustering, image segmentation, image smoothing, feature space, low-level vision.

1 INTRODUCTION

LOW-LEVEL computer vision tasks are misleadingly difficult. Incorrect results can be easily obtained since the employed techniques often rely upon the user correctly guessing the values for the tuning parameters. To improve performance, the execution of low-level tasks should be task driven, i.e., supported by independent high-level information. This approach, however, requires that, first, the low-level stage provides a reliable enough representation of the input and that the feature extraction process be controlled only by very few tuning parameters corresponding to intuitive measures in the input domain.

Feature space-based analysis of images is a paradigm which can achieve the above-stated goals. A feature space is a mapping of the input obtained through the processing of the data in small subsets at a time. For each subset, a parametric representation of the feature of interest is obtained and the result is mapped into a point in the multidimensional space of the parameter. After the entire input is processed, significant features correspond to denser regions in the feature space, i.e., to clusters, and the goal of the analysis is the delineation of these clusters.

The nature of the feature space is application dependent. The subsets employed in the mapping can range from individual pixels, as in the color space representation of an image, to a set of quasi-randomly chosen data points, as in the probabilistic Hough transform. Both the advantage and the disadvantage of the feature space paradigm arise from the global nature of the derived representation of the input. On one hand, all the evidence for the presence of a

significant feature is pooled together, providing excellent tolerance to a noise level which may render local decisions unreliable. On the other hand, features with lesser support in the feature space may not be detected in spite of being salient for the task to be executed. This disadvantage, however, can be largely avoided by either augmenting the feature space with additional (spatial) parameters from the input domain or by robust postprocessing of the input domain guided by the results of the feature space analysis.

Analysis of the feature space is application independent. While there are a plethora of published clustering techniques, most of them are not adequate to analyze feature spaces derived from real data. Methods which rely upon a priori knowledge of the number of clusters present (including those which use optimization of a global criterion to find this number), as well as methods which implicitly assume the same shape (most often elliptical) for all the clusters in the space, are not able to handle the complexity of a real feature space. For a recent survey of such methods, see [29, Section 8].

In Fig. 1, a typical example is shown. The color image in Fig. 1a is mapped into the three-dimensional $L^*u^*v^*$ color space (to be discussed in Section 4). There is a continuous transition between the clusters arising from the dominant colors and a decomposition of the space into elliptical tiles will introduce severe artifacts. Enforcing a Gaussian mixture model over such data is doomed to fail, e.g., [49], and even the use of a robust approach with contaminated Gaussian densities [67] cannot be satisfactory for such complex cases. Note also that the mixture models require the number of clusters as a parameter, which raises its own challenges. For example, the method described in [45] proposes several different ways to determine this number.

Arbitrarily structured feature spaces can be analyzed only by nonparametric methods since these methods do not have embedded assumptions. Numerous nonparametric clustering methods were described in the literature and they can be classified into two large classes: hierarchical clustering and density estimation. Hierarchical clustering techniques either aggregate or divide the data based on

• D. Comaniciu is with the Imaging and Visualization Department, Siemens Corporate Research, 755 College Road East, Princeton, NJ 08540.
E-mail: comanici@scr.siemens.com.

• P. Meer is with the Electrical and Computer Engineering Department, Rutgers University, 94 Brett Road, Piscataway, NJ 08854-8058.
E-mail: meer@caip.rutgers.edu.

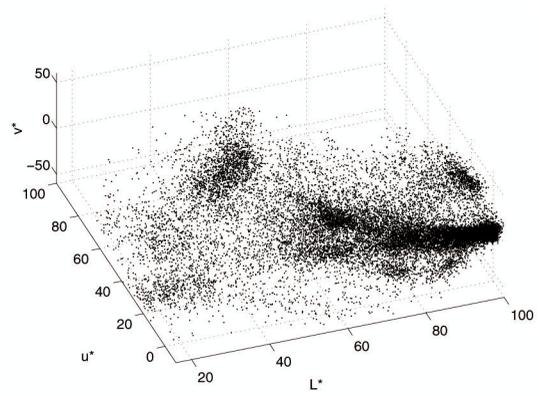
Manuscript received 17 Jan. 2001; revised 16 July 2001; accepted 21 Nov. 2001.

Recommended for acceptance by V. Solo.

For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number 113483.



(a)



(b)

Fig. 1. Example of a feature space. (a) A 400×276 color image. (b) Corresponding $L^*u^*v^*$ color space with 110,400 data points.

some proximity measure. See [28, Section 3.2] for a survey of hierarchical clustering methods. The hierarchical methods tend to be computationally expensive and the definition of a meaningful stopping criterion for the fusion (or division) of the data is not straightforward.

The rationale behind the density estimation-based non-parametric clustering approach is that the feature space can be regarded as the empirical probability density function (p.d.f.) of the represented parameter. Dense regions in the feature space thus correspond to local maxima of the p.d.f., that is, to the *modes* of the unknown density. Once the location of a mode is determined, the cluster associated with it is delineated based on the local structure of the feature space [25], [60], [63].

Our approach to mode detection and clustering is based on the mean shift procedure, proposed in 1975 by Fukunaga and Hostetler [21] and largely forgotten until Cheng's paper [7] rekindled interest in it. In spite of its excellent qualities, the mean shift procedure does not seem to be known in statistical literature. While the book [54, Section 6.2.2] discusses [21], the advantages of employing a mean shift type procedure in density estimation were only recently rediscovered [8].

As will be proven in the sequel, a computational module based on the mean shift procedure is an extremely versatile tool for feature space analysis and can provide reliable solutions for many vision tasks. In Section 2, the mean shift procedure is defined and its properties are analyzed. In Section 3, the procedure is used as the computational module for robust feature space analysis and implementation issues are discussed. In Section 4, the feature space analysis technique is applied to two low-level vision tasks: discontinuity preserving filtering and image segmentation. Both algorithms can have as input either gray level or color images and the only parameter to be tuned by the user is the resolution of the analysis. The applicability of the mean shift procedure is not restricted to the presented examples. In Section 5, other applications are mentioned and the procedure is put into a more general context.

2 THE MEAN SHIFT PROCEDURE

Kernel density estimation (known as the Parzen window technique in pattern recognition literature [17, Section 4.3]) is the most popular density estimation method. Given n data

points \mathbf{x}_i , $i = 1, \dots, n$ in the d -dimensional space R^d , the *multivariate kernel density estimator* with kernel $K(\mathbf{x})$ and a symmetric positive definite $d \times d$ bandwidth matrix \mathbf{H} , computed in the point \mathbf{x} is given by

$$\hat{f}(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^n K_{\mathbf{H}}(\mathbf{x} - \mathbf{x}_i), \quad (1)$$

where

$$K_{\mathbf{H}}(\mathbf{x}) = |\mathbf{H}|^{-1/2} K(\mathbf{H}^{-1/2}\mathbf{x}). \quad (2)$$

The d -variate kernel $K(\mathbf{x})$ is a bounded function with compact support satisfying [62, p. 95]

$$\begin{aligned} \int_{R^d} K(\mathbf{x}) d\mathbf{x} &= 1 & \lim_{\|\mathbf{x}\| \rightarrow \infty} \|\mathbf{x}\|^d K(\mathbf{x}) &= 0 \\ \int_{R^d} \mathbf{x} K(\mathbf{x}) d\mathbf{x} &= 0 & \int_{R^d} \mathbf{x} \mathbf{x}^T K(\mathbf{x}) d\mathbf{x} &= c_K \mathbf{I}, \end{aligned} \quad (3)$$

where c_K is a constant. The multivariate kernel can be generated from a symmetric univariate kernel $K_1(x)$ in two different ways

$$K^P(\mathbf{x}) = \prod_{i=1}^d K_1(x_i) \quad K^S(\mathbf{x}) = a_{k,d} K_1(\|\mathbf{x}\|), \quad (4)$$

where $K^P(\mathbf{x})$ is obtained from the product of the univariate kernels and $K^S(\mathbf{x})$ from rotating $K_1(x)$ in R^d , i.e., $K^S(\mathbf{x})$ is radially symmetric. The constant $a_{k,d}^{-1} = \int_{R^d} K_1(\|\mathbf{x}\|) d\mathbf{x}$ assures that $K^S(\mathbf{x})$ integrates to one, though this condition can be relaxed in our context. Either type of multivariate kernel obeys (3), but, for our purposes, the radially symmetric kernels are often more suitable.

We are interested only in a special class of radially symmetric kernels satisfying

$$K(\mathbf{x}) = c_{k,d} k(\|\mathbf{x}\|^2), \quad (5)$$

in which case it suffices to define the function $k(x)$ called the *profile* of the kernel, only for $x \geq 0$. The normalization constant $c_{k,d}$, which makes $K(\mathbf{x})$ integrate to one, is assumed strictly positive.

Using a fully parameterized \mathbf{H} increases the complexity of the estimation [62, p. 106] and, in practice, the bandwidth matrix \mathbf{H} is chosen either as diagonal $\mathbf{H} = \text{diag}[h_1^2, \dots, h_d^2]$,

or proportional to the identity matrix $\mathbf{H} = h^2 \mathbf{I}$. The clear advantage of the latter case is that only one bandwidth parameter $h > 0$ must be provided; however, as can be seen from (2), then the validity of an Euclidean metric for the feature space should be confirmed first. Employing only one bandwidth parameter, the kernel density estimator (1) becomes the well-known expression

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

The quality of a kernel density estimator is measured by the mean of the square error between the density and its estimate, integrated over the domain of definition. In practice, however, only an asymptotic approximation of this measure (denoted as AMISE) can be computed. Under the asymptotics, the number of data points $n \rightarrow \infty$, while the bandwidth $h \rightarrow 0$ at a rate slower than n^{-1} . For both types of multivariate kernels, the AMISE measure is minimized by the Epanechnikov kernel [51, p. 139], [62, p. 104] having the profile

$$k_E(x) = \begin{cases} 1-x & 0 \leq x \leq 1 \\ 0 & x > 1, \end{cases} \quad (7)$$

which yields the radially symmetric kernel

$$K_E(\mathbf{x}) = \begin{cases} \frac{1}{2} c_d^{-1} (d+2) (1 - \|\mathbf{x}\|^2) & \|\mathbf{x}\| \leq 1 \\ 0 & \text{otherwise,} \end{cases} \quad (8)$$

where c_d is the volume of the unit d -dimensional sphere. Note that the Epanechnikov profile is not differentiable at the boundary. The profile

$$k_N(x) = \exp\left(-\frac{1}{2}x\right) \quad x \geq 0 \quad (9)$$

yields the multivariate normal kernel

$$K_N(\mathbf{x}) = (2\pi)^{-d/2} \exp\left(-\frac{1}{2}\|\mathbf{x}\|^2\right) \quad (10)$$

for both types of composition (4). The normal kernel is often symmetrically truncated to have a kernel with finite support.

While these two kernels will suffice for most applications we are interested in, all the results presented below are valid for arbitrary kernels within the conditions to be stated. Employing the profile notation, the density estimator (6) can be rewritten as

$$\hat{f}_{h,K}(\mathbf{x}) = \frac{c_{k,d}}{nh^d} \sum_{i=1}^n k\left(\left\|\frac{\mathbf{x} - \mathbf{x}_i}{h}\right\|^2\right). \quad (11)$$

The first step in the analysis of a feature space with the underlying density $f(\mathbf{x})$ is to find the modes of this density. The modes are located among the zeros of the gradient $\nabla f(\mathbf{x}) = \mathbf{0}$ and the mean shift procedure is an elegant way to locate these zeros *without* estimating the density.

2.1 Density Gradient Estimation

The density gradient estimator is obtained as the gradient of the density estimator by exploiting the linearity of (11)

$$\hat{\nabla} f_{h,K}(\mathbf{x}) \equiv \nabla \hat{f}_{h,K}(\mathbf{x}) = \frac{2c_{k,d}}{nh^{d+2}} \sum_{i=1}^n (\mathbf{x} - \mathbf{x}_i) k'\left(\left\|\frac{\mathbf{x} - \mathbf{x}_i}{h}\right\|^2\right). \quad (12)$$

We define the function

$$g(x) = -k'(x), \quad (13)$$

assuming that the derivative of the kernel profile k exists for all $x \in [0, \infty)$, except for a finite set of points. Now, using $g(x)$ for profile, the kernel $G(\mathbf{x})$ is defined as

$$G(\mathbf{x}) = c_{g,d} g\left(\|\mathbf{x}\|^2\right), \quad (14)$$

where $c_{g,d}$ is the corresponding normalization constant. The kernel $K(\mathbf{x})$ was called the shadow of $G(\mathbf{x})$ in [7] in a slightly different context. Note that the Epanechnikov kernel is the shadow of the uniform kernel, i.e., the d -dimensional unit sphere, while the normal kernel and its shadow have the same expression.

Introducing $g(x)$ into (12) yields,

$$\begin{aligned} \hat{\nabla} f_{h,K}(\mathbf{x}) &= \frac{2c_{k,d}}{nh^{d+2}} \sum_{i=1}^n (\mathbf{x}_i - \mathbf{x}) g\left(\left\|\frac{\mathbf{x} - \mathbf{x}_i}{h}\right\|^2\right) \\ &= \frac{2c_{k,d}}{nh^{d+2}} \left[\sum_{i=1}^n g\left(\left\|\frac{\mathbf{x} - \mathbf{x}_i}{h}\right\|^2\right) \right] \left[\frac{\sum_{i=1}^n \mathbf{x}_i g\left(\left\|\frac{\mathbf{x} - \mathbf{x}_i}{h}\right\|^2\right)}{\sum_{i=1}^n g\left(\left\|\frac{\mathbf{x} - \mathbf{x}_i}{h}\right\|^2\right)} - \mathbf{x} \right], \end{aligned} \quad (15)$$

where $\sum_{i=1}^n g\left(\left\|\frac{\mathbf{x} - \mathbf{x}_i}{h}\right\|^2\right)$ is assumed to be a positive number. This condition is easy to satisfy for all the profiles met in practice. Both terms of the product in (15) have special significance. From (11), the first term is proportional to the density estimate at \mathbf{x} computed with the kernel G

$$\hat{f}_{h,G}(\mathbf{x}) = \frac{c_{g,d}}{nh^d} \sum_{i=1}^n g\left(\left\|\frac{\mathbf{x} - \mathbf{x}_i}{h}\right\|^2\right). \quad (16)$$

The second term is the *mean shift*

$$\mathbf{m}_{h,G}(\mathbf{x}) = \frac{\sum_{i=1}^n \mathbf{x}_i g\left(\left\|\frac{\mathbf{x} - \mathbf{x}_i}{h}\right\|^2\right)}{\sum_{i=1}^n g\left(\left\|\frac{\mathbf{x} - \mathbf{x}_i}{h}\right\|^2\right)} - \mathbf{x}, \quad (17)$$

i.e., the difference between the weighted mean, using the kernel G for weights, and \mathbf{x} , the center of the kernel (window). From (16) and (17), (15) becomes

$$\hat{\nabla} f_{h,K}(\mathbf{x}) = \hat{f}_{h,G}(\mathbf{x}) \frac{2c_{k,d}}{h^2 c_{g,d}} \mathbf{m}_{h,G}(\mathbf{x}), \quad (18)$$

yielding

$$\mathbf{m}_{h,G}(\mathbf{x}) = \frac{1}{2} h^2 c \frac{\hat{\nabla} f_{h,K}(\mathbf{x})}{\hat{f}_{h,G}(\mathbf{x})}. \quad (19)$$

The expression (19) shows that, at location \mathbf{x} , the mean shift vector computed with kernel G is proportional to the *normalized* density gradient estimate obtained with kernel K . The normalization is by the density estimate in \mathbf{x} computed with the kernel G . The mean shift vector thus always points toward the direction of maximum increase in the density.

This is a more general formulation of the property first remarked by Fukunaga and Hostetler [20, p. 535], [21], and discussed in [7].

The relation captured in (19) is intuitive, the local mean is shifted toward the region in which the majority of the points reside. Since the mean shift vector is aligned with the local gradient estimate, it can define a path leading to a stationary point of the *estimated* density. The modes of the density are such stationary points. The *mean shift procedure*, obtained by successive

- computation of the mean shift vector $\mathbf{m}_{h,G}(\mathbf{x})$,
- translation of the kernel (window) $G(\mathbf{x})$ by $\mathbf{m}_{h,G}(\mathbf{x})$,

is guaranteed to converge at a nearby point where the estimate (11) has zero gradient, as will be shown in the next section. The presence of the normalization by the density estimate is a desirable feature. The regions of low-density values are of no interest for the feature space analysis and, in such regions, the mean shift steps are large. Similarly, near local maxima the steps are small and the analysis more refined. The mean shift procedure thus is an adaptive gradient ascent method.

2.2 Sufficient Condition for Convergence

Denote by $\{\mathbf{y}_j\}_{j=1,2,\dots}$ the sequence of successive locations of the kernel G , where, from (17),

$$\mathbf{y}_{j+1} = \frac{\sum_{i=1}^n \mathbf{x}_i g\left(\left\|\frac{\mathbf{x}-\mathbf{x}_i}{h}\right\|^2\right)}{\sum_{i=1}^n g\left(\left\|\frac{\mathbf{x}-\mathbf{x}_i}{h}\right\|^2\right)} \quad j = 1, 2, \dots \quad (20)$$

is the weighted mean at \mathbf{y}_j computed with kernel G and \mathbf{y}_1 is the center of the initial position of the kernel. The corresponding sequence of density estimates computed with kernel K , $\{\hat{f}_{h,K}(j)\}_{j=1,2,\dots}$ is given by

$$\hat{f}_{h,K}(j) = \hat{f}_{h,K}(\mathbf{y}_j) \quad j = 1, 2, \dots \quad (21)$$

As stated by the following theorem, a kernel K that obeys some mild conditions suffices for the convergence of the sequences $\{\mathbf{y}_j\}_{j=1,2,\dots}$ and $\{\hat{f}_{h,K}(j)\}_{j=1,2,\dots}$.

Theorem 1. *If the kernel K has a convex and monotonically decreasing profile, the sequences $\{\mathbf{y}_j\}_{j=1,2,\dots}$ and $\{\hat{f}_{h,K}(j)\}_{j=1,2,\dots}$ converge and $\{\hat{f}_{h,K}(j)\}_{j=1,2,\dots}$ is monotonically increasing.*

The proof is given in the Appendix. The theorem generalizes the result derived differently in [13], where K was the Epanechnikov kernel and G the uniform kernel. The theorem remains valid when each data point \mathbf{x}_i is associated with a nonnegative weight w_i . An example of nonconvergence when the kernel K is not convex is shown in [10, p. 16].

The convergence property of the mean shift was also discussed in [7, Section iv]. (Note, however, that almost all the discussion there is concerned with the “blurring” process in which the input is recursively modified after each mean shift step.) The convergence of the procedure as defined in this paper was attributed in [7] to the gradient ascent nature of (19). However, as shown in [4, Section 1.2], moving in the direction of the local gradient guarantees convergence only for infinitesimal steps. The step size of a gradient-based algorithm is crucial for the overall performance. If the step size is too large, the algorithm will diverge, while if the step size is too small, the rate of convergence may be very slow. A number of

costly procedures have been developed for step size selection [4, p. 24]. The guaranteed convergence (as shown by Theorem 1) is due to the adaptive magnitude of the mean shift vector, which also eliminates the need for additional procedures to choose the adequate step sizes. This is a major advantage over the traditional gradient-based methods.

For discrete data, the number of steps to convergence depends on the employed kernel. When G is the uniform kernel, convergence is achieved in a finite number of steps since the number of locations generating distinct mean values is finite. However, when the kernel G imposes a weighting on the data points (according to the distance from its center), the mean shift procedure is infinitely convergent. The practical way to stop the iterations is to set a lower bound for the magnitude of the mean shift vector.

2.3 Mean Shift-Based Mode Detection

Let us denote by \mathbf{y}_c and $\hat{f}_{h,K}^c = \hat{f}_{h,K}(\mathbf{y}_c)$ the convergence points of the sequences $\{\mathbf{y}_j\}_{j=1,2,\dots}$ and $\{\hat{f}_{h,K}(j)\}_{j=1,2,\dots}$, respectively. The implications of Theorem 1 are the following.

First, the magnitude of the mean shift vector converges to zero. Indeed, from (17) and (20) the j th mean shift vector is

$$\mathbf{m}_{h,G}(\mathbf{y}_j) = \mathbf{y}_{j+1} - \mathbf{y}_j \quad (22)$$

and, at the limit, $\mathbf{m}_{h,G}(\mathbf{y}_c) = \mathbf{y}_c - \mathbf{y}_c = \mathbf{0}$. In other words, the gradient of the density estimate (11) computed at \mathbf{y}_c is zero

$$\nabla \hat{f}_{h,K}(\mathbf{y}_c) = 0, \quad (23)$$

due to (19). Hence, \mathbf{y}_c is a stationary point of $\hat{f}_{h,K}$. Second, since $\{\hat{f}_{h,K}(j)\}_{j=1,2,\dots}$ is monotonically increasing, the mean shift iterations satisfy the conditions required by the *Capture Theorem* [4, p. 45], which states that the trajectories of such gradient methods are attracted by local maxima if they are unique (within a small neighborhood) stationary points. That is, once \mathbf{y}_j gets sufficiently close to a mode of $\hat{f}_{h,K}$, it converges to it. The set of all locations that converge to the same mode defines the *basin of attraction* of that mode.

The theoretical observations from above suggest a practical algorithm for mode detection:

- Run the mean shift procedure to find the stationary points of $\hat{f}_{h,K}$,
- Prune these points by retaining only the local maxima.

The local maxima points are defined, according to the Capture Theorem, as unique stationary points within some small open sphere. This property can be tested by perturbing each stationary point by a random vector of small norm and letting the mean shift procedure converge again. Should the point of convergence be unchanged (up to a tolerance), the point is a local maximum.

2.4 Smooth Trajectory Property

The mean shift procedure employing a normal kernel has an interesting property. Its path toward the mode follows a smooth trajectory, the angle between two consecutive mean shift vectors being always less than 90 degrees.

Using the normal kernel (10), the j th mean shift vector is given by

$$\mathbf{m}_{h,N}(\mathbf{y}_j) = \mathbf{y}_{j+1} - \mathbf{y}_j = \frac{\sum_{i=1}^n \mathbf{x}_i \exp\left(\left\|\frac{\mathbf{x}-\mathbf{x}_i}{h}\right\|^2\right)}{\sum_{i=1}^n \exp\left(\left\|\frac{\mathbf{x}-\mathbf{x}_i}{h}\right\|^2\right)} - \mathbf{y}_j. \quad (24)$$

The following theorem holds true for all $j = 1, 2, \dots$, according to the proof given in the Appendix.

Theorem 2. *The cosine of the angle between two consecutive mean shift vectors is strictly positive when a normal kernel is employed, i.e.,*

$$\frac{\mathbf{m}_{h,N}(\mathbf{y}_j)^\top \mathbf{m}_{h,N}(\mathbf{y}_{j+1})}{\|\mathbf{m}_{h,N}(\mathbf{y}_j)\| \|\mathbf{m}_{h,N}(\mathbf{y}_{j+1})\|} > 0. \quad (25)$$

As a consequence of Theorem 2, the normal kernel appears to be the optimal one for the mean shift procedure. The smooth trajectory of the mean shift procedure is in contrast with the standard steepest ascent method [4, p. 21] (local gradient evaluation followed by line maximization) whose convergence rate on surfaces with deep narrow valleys is slow due to its zigzagging trajectory.

In practice, the convergence of the mean shift procedure based on the normal kernel requires large number of steps, as was discussed at the end of Section 2.2. Therefore, in most of our experiments, we have used the uniform kernel, for which the convergence is finite, and not the normal kernel. Note, however, that the quality of the results almost always improves when the normal kernel is employed.

2.5 Relation to Kernel Regression

Important insight can be gained when (19) is obtained approaching the problem differently. Considering the univariate case suffices for this purpose.

Kernel regression is a nonparametric method to estimate complex trends from noisy data. See [62, chapter 5] for an introduction to the topic, [24] for a more in-depth treatment. Let n measured data points be (X_i, Z_i) and assume that the values X_i are the outcomes of a random variable x with probability density function $f(x)$, $x_i = X_i$, $i = 1, \dots, n$, while the relation between Z_i and X_i is

$$Z_i = m(X_i) + \epsilon_i \quad i = 1, \dots, n, \quad (26)$$

where $m(x)$ is called the regression function and ϵ_i is an independently distributed, zero-mean error, $E[\epsilon_i] = 0$.

A natural way to estimate the regression function is by locally fitting a degree p polynomial to the data. For a window centered at x , the polynomial coefficients then can be obtained by weighted least squares, the weights being computed from a symmetric function $g(x)$. The size of the window is controlled by the parameter h , $g_h(x) = h^{-1}g(x/h)$. The simplest case is that of fitting a constant to the data in the window, i.e., $p = 0$. It can be shown, [24, Section 3.1], [62, Section 5.2], that the estimated constant is the value of the *Nadaraya-Watson* estimator,

$$\hat{m}(x; h) = \frac{\sum_{i=1}^n g_h(x - X_i) Z_i}{\sum_{i=1}^n g_h(x - X_i)}, \quad (27)$$

introduced in the statistical literature 35 years ago. The asymptotic conditional bias of the estimator has the expression [24, p. 109], [62, p. 125],

$$\begin{aligned} E[(\hat{m}(x; h) - m(x)) \mid X_1, \dots, X_n] \\ \approx h^2 \frac{m''(x)f(x) + 2m'(x)f'(x)}{2f(x)} \mu_2[g], \end{aligned} \quad (28)$$

where $\mu_2[g] = \int u^2 g(u) du$. Defining $m(x) = x$ reduces the Nadaraya-Watson estimator to (20) (in the univariate case), while (28) becomes

$$E[(\hat{x} - x) \mid X_1, \dots, X_n] \approx h^2 \frac{f'(x)}{f(x)\mu_2[g]}, \quad (29)$$

which is similar to (19). The mean shift procedure thus exploits to its advantage the inherent bias of the zero-order kernel regression.

The connection to the kernel regression literature opens many interesting issues, however, most of these are more of a theoretical than practical importance.

2.6 Relation to Location M-Estimators

The M-estimators are a family of robust techniques which can handle data in the presence of severe contaminations, i.e., outliers. See [26], [32] for introductory surveys. In our context only, the problem of location estimation has to be considered.

Given the data \mathbf{x}_i , $i = 1, \dots, n$, and the scale h , will define $\hat{\theta}$, the location estimator as

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} J(\theta) = \underset{\theta}{\operatorname{argmin}} \sum_{i=1}^n \rho\left(\left\|\frac{\theta - \mathbf{x}_i}{h}\right\|^2\right), \quad (30)$$

where, $\rho(u)$ is a symmetric, nonnegative valued function, with a unique minimum at the origin and nondecreasing for $u \geq 0$. The estimator is obtained from the normal equations

$$\nabla_{\theta} J(\hat{\theta}) = 2h^{-2}(\hat{\theta} - \mathbf{x}_i)w\left(\left\|\frac{\hat{\theta} - \mathbf{x}_i}{h}\right\|^2\right) = \mathbf{0}, \quad (31)$$

where

$$w(u) = \frac{d\rho(u)}{du}.$$

Therefore, the iterations to find the location M-estimate are based on

$$\hat{\theta} = \frac{\sum_{i=1}^n \mathbf{x}_i w\left(\left\|\frac{\hat{\theta} - \mathbf{x}_i}{h}\right\|^2\right)}{\sum_{i=1}^n w\left(\left\|\frac{\hat{\theta} - \mathbf{x}_i}{h}\right\|^2\right)}, \quad (32)$$

which is identical to (20) when $w(u) \equiv g(u)$. Taking into account (13), the minimization (30) becomes

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^n k\left(\left\|\frac{\theta - \mathbf{x}_i}{h}\right\|^2\right), \quad (33)$$

which can also be interpreted as

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} \hat{f}_{h,K}(\theta \mid \mathbf{x}_1, \dots, \mathbf{x}_n). \quad (34)$$

That is, the location estimator is the mode of the density estimated with the kernel K from the available data. Note that the convexity of the $k(x)$ profile, the sufficient condition for the convergence of the mean shift procedure (Section 2.2) is in

accordance with the requirements to be satisfied by the objective function $\rho(u)$.

The relation between location M-estimators and kernel density estimation is not well-investigated in the statistical literature, only [9] discusses it in the context of an edge preserving smoothing technique.

3 ROBUST ANALYSIS OF FEATURE SPACES

Multimodality and arbitrarily shaped clusters are the defining properties of a real feature space. The quality of the mean shift procedure to move toward the mode (peak) of the hill on which it was initiated makes it the ideal computational module to analyze such spaces. To detect all the significant modes, the basic algorithm given in Section 2.3 should be run multiple times (evolving in principle in parallel) with initializations that cover the entire feature space.

Before the analysis is performed, two important (and somewhat related) issues should be addressed: the metric of the feature space and the shape of the kernel. The mapping from the input domain into a feature space often associates a non-Euclidean metric to the space. The problem of color representation will be discussed in Section 4, but the employed parameterization has to be carefully examined even in a simple case like the Hough space of lines, e.g., [48], [61].

The presence of a Mahalanobis metric can be accommodated by an adequate choice of the bandwidth matrix (2). In practice, however, it is preferable to have assured that the metric of the feature space is Euclidean and, thus, the bandwidth matrix is controlled by a single parameter, $\mathbf{H} = h^2\mathbf{I}$. To be able to use the same kernel size for all the mean shift procedures in the feature space, the necessary condition is that local density variations near a significant mode are not as large as the entire support of a significant mode somewhere else.

The starting points of the mean shift procedures should be chosen to have the entire feature space (except the very sparse regions) tessellated by the kernels (windows). Regular tessellations are not required. As the windows evolve toward the modes, almost all the data points are visited and, thus, all the information captured in the feature space is exploited. Note that the convergence to a given mode may yield slightly different locations due to the threshold that terminates the iterations. Similarly, on flat plateaus, the value of the gradient is close to zero and the mean shift procedure could stop.

These artifacts are easy to eliminate through postprocessing. Mode candidates at a distance less than the kernel bandwidth are fused, the one corresponding to the highest density being chosen. The global structure of the feature space can be confirmed by measuring the significance of the valleys defined along a cut through the density in the direction determined by two modes.

The delineation of the clusters is a natural outcome of the mode seeking process. After convergence, the *basin of attraction* of a mode, i.e., the data points visited by *all* the mean shift procedures converging to that mode, automatically delineates a cluster of arbitrary shape. Close to the boundaries, where a data point could have been visited by several diverging procedures, majority logic can be employed. It is important to notice that, in computer vision, most often we are not dealing with an abstract clustering

problem. The input domain almost always provides an independent test for the validity of *local decisions* in the feature space. That is, while it is less likely that one can recover from a severe clustering error, allocation of a few uncertain data points can be reliably supported by input domain information.

The multimodal feature space analysis technique was discussed in detail in [12]. It was shown experimentally, that for a synthetic, bimodal normal distribution, the technique achieves a classification error similar to the optimal Bayesian classifier. The behavior of this feature space analysis technique is illustrated in Fig. 2. A two-dimensional data set of 110,400 points (Fig. 2a) is decomposed into seven clusters represented with different colors in Fig. 2b. A number of 159 mean shift procedures with uniform kernel were employed. Their trajectories are shown in Fig. 2c, overlapped over the density estimate computed with the Epanechnikov kernel. The pruning of the mode candidates produced seven peaks. Observe that some of the trajectories are prematurely stopped by local plateaus.

3.1 Bandwidth Selection

The influence of the bandwidth parameter h was assessed empirically in [12] through a simple image segmentation task. In a more rigorous approach, however, four different techniques for bandwidth selection can be considered.

- The first one has a statistical motivation. The optimal bandwidth associated with the kernel density estimator (6) is defined as the bandwidth that achieves the best compromise between the bias and variance of the estimator, over all $\mathbf{x} \in R^d$, i.e., minimizes AMISE. In the multivariate case, the resulting bandwidth formula [54, p. 85], [62, p. 99] is of little practical use, since it depends on the Laplacian of the unknown density being estimated, and its performance is not well understood [62, p. 108]. For the univariate case, a reliable method for bandwidth selection is the plug-in rule [53], which was proven to be superior to least-squares cross-validation and biased cross-validation [42], [55, p. 46]. Its only assumption is the smoothness of the underlying density.
- The second bandwidth selection technique is related to the stability of the decomposition. The bandwidth is taken as the center of the largest operating range over which the same number of clusters are obtained for the given data [20, p. 541].
- For the third technique, the best bandwidth maximizes an objective function that expresses the quality of the decomposition (i.e., the index of cluster validity). The objective function typically compares the inter- versus intra-cluster variability [30], [28] or evaluates the isolation and connectivity of the delineated clusters [43].
- Finally, since in most of the cases the decomposition is task dependent, top-down information provided by the user or by an upper-level module can be used to control the kernel bandwidth.

We present in [15], a detailed analysis of the bandwidth selection problem. To solve the difficulties generated by the narrow peaks and the tails of the underlying density, two locally adaptive solutions are proposed. One is nonparametric, being based on a newly defined adaptive mean shift

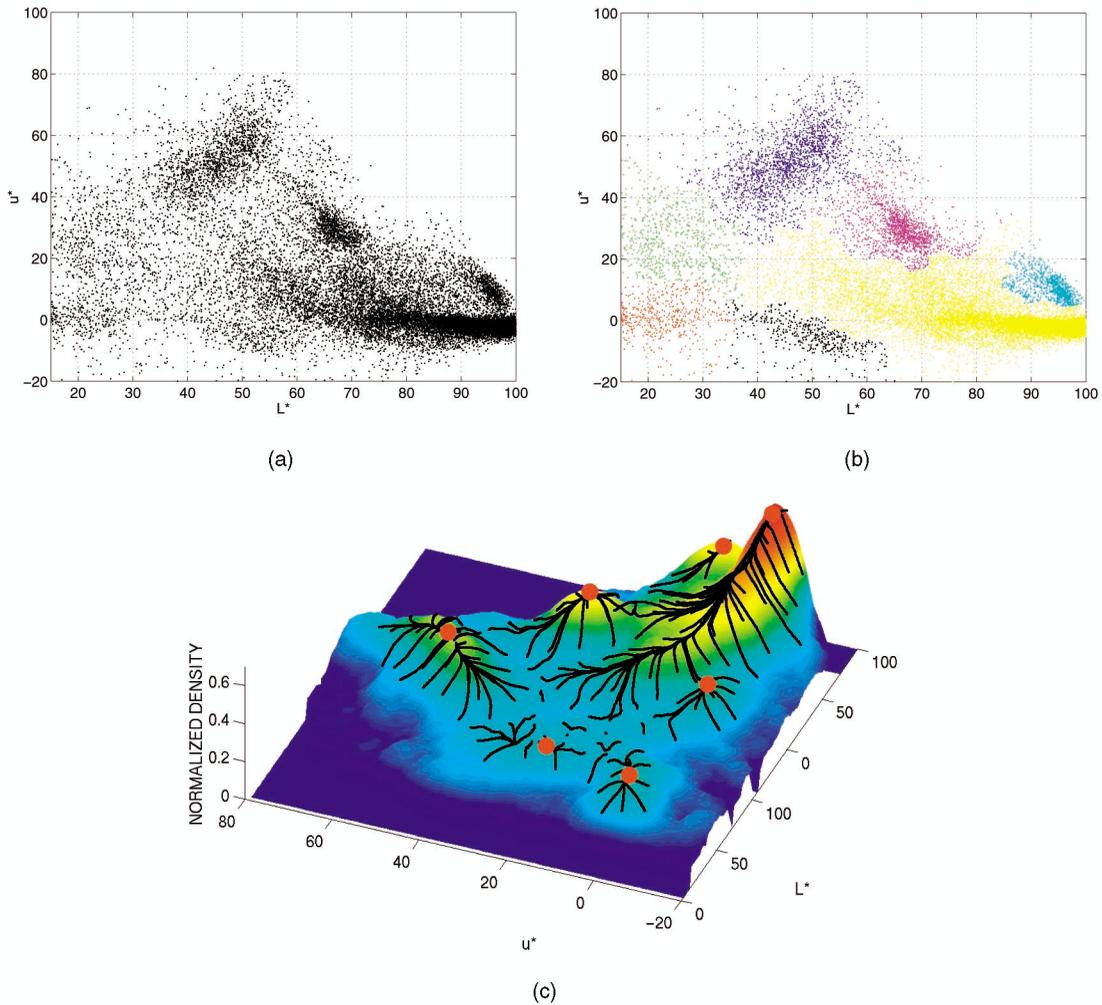


Fig. 2. Example of a 2D feature space analysis. (a) Two-dimensional data set of 110,400 points representing the first two components of the $L^*u^*v^*$ space shown in Fig. 1b. (b) Decomposition obtained by running 159 mean shift procedures with different initializations. (c) Trajectories of the mean shift procedures drawn over the Epanechnikov density estimate computed for the same data set. The peaks retained for the final classification are marked with red dots.

procedure, which exploits the plug-in rule and the sample point density estimator. The other is semiparametric, imposing a local structure on the data to extract reliable scale information. We show that the local bandwidth should maximize the magnitude of the normalized mean shift vector. The adaptation of the bandwidth provides superior results when compared to the fixed bandwidth procedure. For more details, see [15].

3.2 Implementation Issues

An efficient computation of the mean shift procedure first requires the resampling of the input data with a regular grid. This is a standard technique in the context of density estimation which leads to a *binned estimator* [62, Appendix D]. The procedure is similar to defining a histogram where linear interpolation is used to compute the weights associated with the grid points. Further reduction in the computation time is achieved by employing algorithms for multidimensional range searching [52, p. 373] used to find the data points falling in the neighborhood of a given kernel. For the efficient Euclidean distance computation, we used the improved absolute error inequality criterion, derived in [39].

4 APPLICATIONS

The feature space analysis technique introduced in the previous section is application independent and, thus, can be used to develop vision algorithms for a wide variety of tasks. Two somewhat related applications are discussed in the sequel: discontinuity preserving smoothing and image segmentation. The versatility of the feature space analysis enables the design of algorithms in which the user controls performance through a single parameter, the resolution of the analysis (i.e., bandwidth of the kernel). Since the control parameter has clear physical meaning, the new algorithms can be easily integrated into systems performing more complex tasks. Furthermore, both gray level and color images are processed with the same algorithm, in the former case, the feature space containing two degenerate dimensions that have no effect on the mean shift procedure.

Before proceeding to develop the new algorithms, the issue of the employed color space has to be settled. To obtain a meaningful segmentation, *perceived* color differences should correspond to Euclidean distances in the color space chosen to represent the features (pixels). An Euclidean metric, however, is not guaranteed for a color space [65, Sections 6.5.2, 8.4]. The spaces $L^*u^*v^*$ and $L^*a^*b^*$ were especially designed to best

approximate perceptually uniform color spaces. In both cases, L^* , the *lightness* (relative brightness) coordinate, is defined the same way, the two spaces differ only through the chromaticity coordinates. The dependence of all three coordinates on the traditional *RGB* color values is nonlinear. See [46, Section 3.5] for a readily accessible source for the conversion formulae. The metric of perceptually uniform color spaces is discussed in the context of feature representation for image segmentation in [16]. In practice, there is no clear advantage between using $L^*u^*v^*$ or $L^*a^*b^*$; in the proposed algorithms, we employed $L^*u^*v^*$ motivated by a linear mapping property [65, p.166].

Our first image segmentation algorithm was a straightforward application of the feature space analysis technique to an $L^*u^*v^*$ representation of the color image [11]. The modularity of the segmentation algorithm enabled its integration by other groups to a large variety of applications like image retrieval [1], face tracking [6], object-based video coding for MPEG-4 [22], shape detection and recognition [33], and texture analysis [47], to mention only a few. However, since the feature space analysis can be applied unchanged to moderately higher dimensional spaces (see Section 5), we subsequently also incorporated the spatial coordinates of a pixel into its feature space representation. This *joint domain* representation is employed in the two algorithms described here.

An image is typically represented as a two-dimensional lattice of p -dimensional vectors (pixels), where $p = 1$ in the gray-level case, three for color images, and $p > 3$ in the multispectral case. The space of the lattice is known as the *spatial* domain, while the gray level, color, or spectral information is represented in the *range* domain. For both domains, Euclidean metric is assumed. When the location and range vectors are concatenated in the joint spatial-range domain of dimension $d = p + 2$, their different nature has to be compensated by proper normalization. Thus, the multivariate kernel is defined as the product of two radially symmetric kernels and the Euclidean metric allows a single bandwidth parameter for each domain

$$K_{h_s, h_r}(\mathbf{x}) = \frac{C}{h_s^2 h_r^p} k\left(\left\|\frac{\mathbf{x}^s}{h_s}\right\|^2\right) k\left(\left\|\frac{\mathbf{x}^r}{h_r}\right\|^2\right), \quad (35)$$

where \mathbf{x}^s is the spatial part, \mathbf{x}^r is the range part of a feature vector, $k(x)$ the common profile used in both two domains, h_s and h_r the employed kernel bandwidths, and C the corresponding normalization constant. In practice, an Epanechnikov or a (truncated) normal kernel always provides satisfactory performance, so the user only has to set the bandwidth parameter $\mathbf{h} = (h_s, h_r)$, which, by controlling the size of the kernel, determines the resolution of the mode detection.

4.1 Discontinuity Preserving Smoothing

Smoothing through replacing the pixel in the center of a window by the (weighted) average of the pixels in the window indiscriminately blurs the image, removing not only the noise but also salient information. Discontinuity preserving smoothing techniques, on the other hand, adaptively reduce the amount of smoothing near abrupt changes in the local structure, i.e., edges.

There are a large variety of approaches to achieve this goal, from adaptive Wiener filtering [31], to implementing isotropic [50] and anisotropic [44] local diffusion processes, a topic which recently received renewed interest [19], [37], [56]. The diffusion-based techniques, however, do not have a straightforward stopping criterion and, after a sufficiently large number of iterations, the processed image collapses into a flat surface. The connection between anisotropic diffusion and M-estimators is analyzed in [5].

A recently proposed noniterative discontinuity preserving smoothing technique is the bilateral filtering [59]. The relation between bilateral filtering and diffusion-based techniques was analyzed in [3]. The bilateral filters also work in the joint spatial-range domain. The data is independently weighted in the two domains and the center pixel is computed as the weighted average of the window. The fundamental difference between the bilateral filtering and the mean shift-based smoothing algorithm is in the use of local information.

4.1.1 Mean Shift Filtering

Let \mathbf{x}_i and $\mathbf{z}_i, i = 1, \dots, n$, be the d -dimensional input and filtered image pixels in the joint spatial-range domain. For each pixel,

1. Initialize $j = 1$ and $\mathbf{y}_{i,1} = \mathbf{x}_i$.
2. Compute $\mathbf{y}_{i,j+1}$ according to (20) until convergence, $\mathbf{y} = \mathbf{y}_{i,c}$.
3. Assign $\mathbf{z}_i = (\mathbf{x}_i^s, \mathbf{y}_{i,c}^r)$.

The superscripts s and r denote the spatial and range components of a vector, respectively. The assignment specifies that the filtered data at the spatial location \mathbf{x}_i^s will have the range component of the point of convergence $\mathbf{y}_{i,c}^r$.

The kernel (window) in the mean shift procedure moves in the direction of the maximum increase in the *joint* density gradient, while the bilateral filtering uses a fixed, static window. In the image smoothed by mean shift filtering, information *beyond* the individual windows is also taken into account.

An important connection between filtering in the joint domain and robust M-estimation should be mentioned. The improved performance of the generalized M-estimators (GM or bounded-influence estimators) is due to the presence of a second weight function which offsets the influence of leverage points, i.e., outliers in the input domain [32, Section 8E]. A similar (at least in spirit) twofold weighting is employed in the bilateral and mean shift-based filterings, which is the main reason for their excellent smoothing performance.

Mean shift filtering with uniform kernel having $(h_s, h_r) = (8, 4)$ has been applied to the often used 256×256 gray-level *cameraman* image (Fig. 3a), the result being shown in Fig. 3b. The regions containing the grass field have been almost completely smoothed, while details such as the tripod and the buildings in the background were preserved. The processing required fractions of a second on a standard PC (600 Mhz Pentium III) using an optimized C++ implementation of the algorithm. On the average, 3.06 iterations were necessary until the filtered value of a pixel was defined, i.e., its mean shift procedure converged.

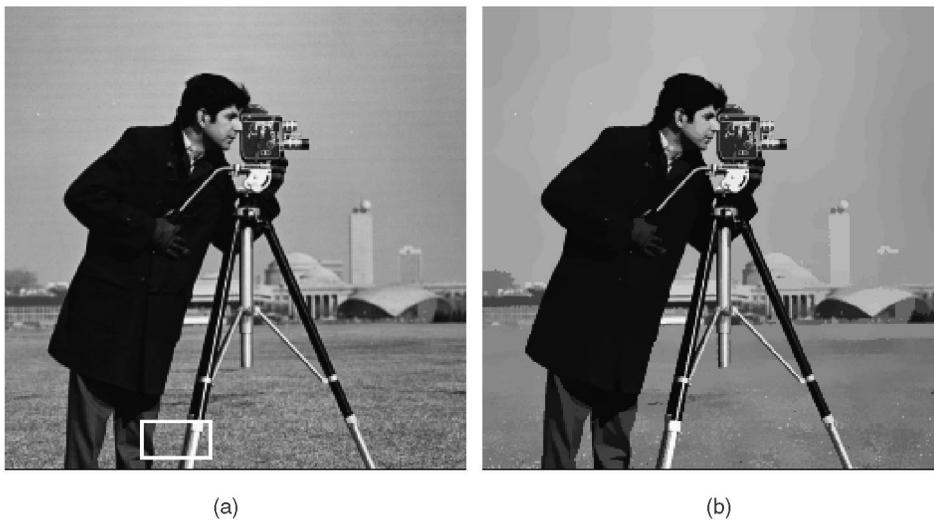


Fig. 3. *Cameraman* image. (a) Original. (b) Mean shift filtered $(h_s, h_r) = (8, 4)$.

To better visualize the filtering process, the 40×20 window marked in Fig. 3a is represented in three dimensions in Fig. 4a. Note that the data was reflected over the horizontal axis of the window for a more informative display. In Fig. 4b, the mean shift paths associated with every other pixel (in both coordinates) from the plateau and the line are shown. Note that convergence points (black dots) are situated in the center of the plateau, away from the discontinuities delineating it. Similarly, the mean shift trajectories on the line remain on it. As a result, the filtered data (Fig. 4c) shows clean quasi-homogeneous regions.

The physical interpretation of the mean shift-based filtering is easy to see by examining Fig. 4a, which, in fact,

displays the three dimensions of the joint domain of a gray-level image. Take a pixel on the line. The uniform kernel defines a parallelepiped centered on this pixel and the computation of the mean shift vector takes into account only those pixels which have *both* their spatial coordinates *and* gray-level values inside the parallelepiped. Thus, if the parallelepiped is not too large, only pixels on the line are averaged and the new location of the window is guaranteed to remain on it.

A second filtering example is shown in Fig. 5. The 512×512 color image *baboon* was processed with mean shift filters employing normal kernels defined using various spatial and range resolutions, $(h_s, h_r) = (8 \div 32, 4 \div 16)$.

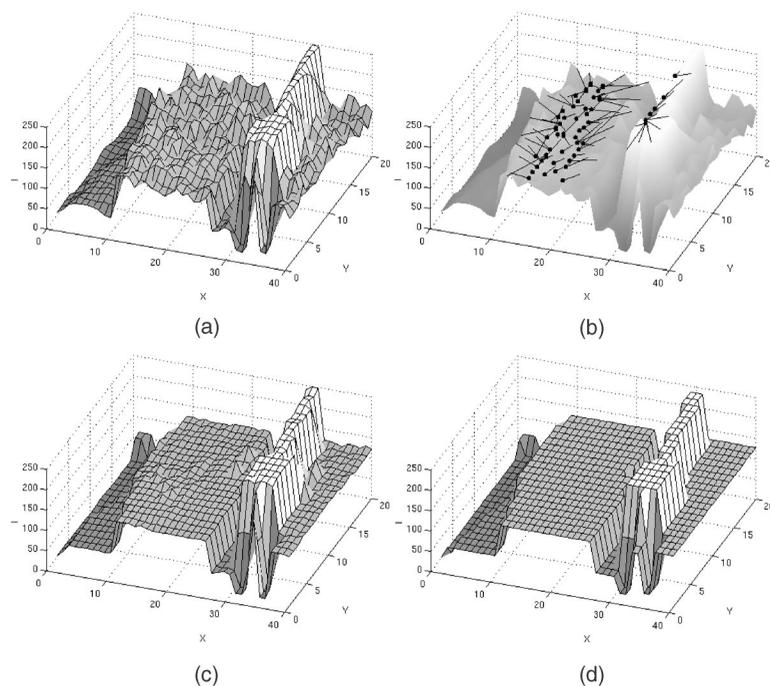


Fig. 4. Visualization of mean shift-based filtering and segmentation for gray-level data. (a) Input. (b) Mean shift paths for the pixels on the plateau and on the line. The black dots are the points of convergence. (c) Filtering result $(h_s, h_r) = (8, 4)$. (d) Segmentation result.

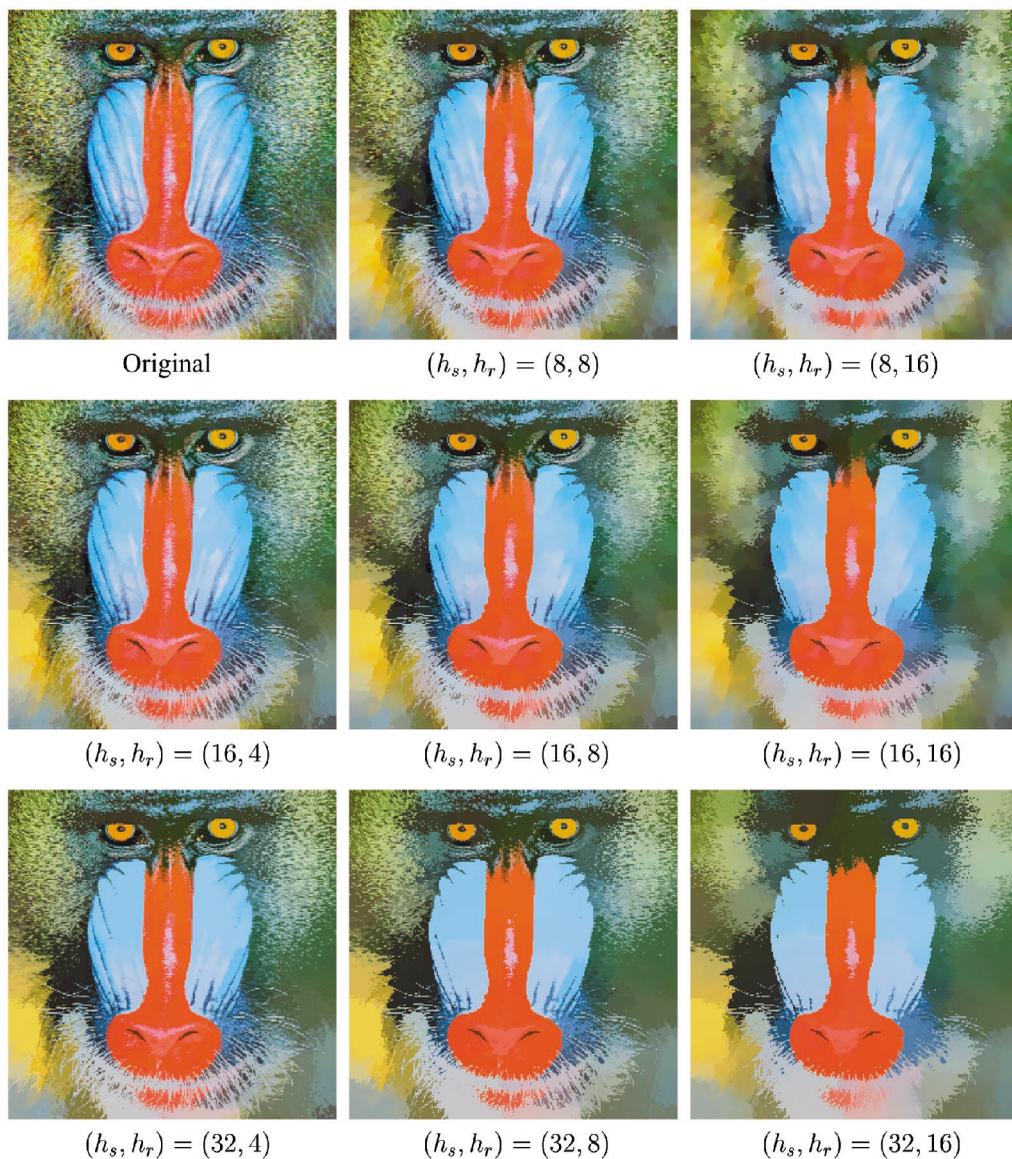


Fig. 5. *Baboon* image. Original and filtered.

While the texture of the fur has been removed, the details of the eyes and the whiskers remained crisp (up to a certain resolution). One can see that the spatial bandwidth has a distinct effect on the output when compared to the range (color) bandwidth. Only features with large spatial support are represented in the filtered image when h_s increases. On the other hand, only features with high color contrast survive when h_r is large. Similar behavior was also reported for the bilateral filter [59, Fig. 3].

4.2 Image Segmentation

Image segmentation, decomposition of a gray level or color image into homogeneous tiles, is arguably the most important low-level vision task. Homogeneity is usually defined as similarity in pixel values, i.e., a piecewise constant model is enforced over the image. From the diversity of image segmentation methods proposed in the literature, we will mention only some whose basic processing relies on the joint

domain. In each case, a vector field is defined over the sampling lattice of the image.

The attraction force field defined in [57] is computed at each pixel as a vector sum of pairwise affinities between the current pixel and all other pixels, with similarity measured in both spatial and range domains. The region boundaries are then identified as loci where the force vectors diverge. It is interesting to note that, for a given pixel, the magnitude and orientation of the force field are similar to those of the joint domain mean shift vector computed at that pixel and projected into the spatial domain. However, in contrast to [57], the mean shift procedure moves in the direction of this vector, away from the boundaries.

The edge flow in [34] is obtained at each location for a given set of directions as the magnitude of the gradient of a smoothed image. The boundaries are detected at image locations which encounter two opposite directions of flow. The quantization of the edge flow direction, however, may

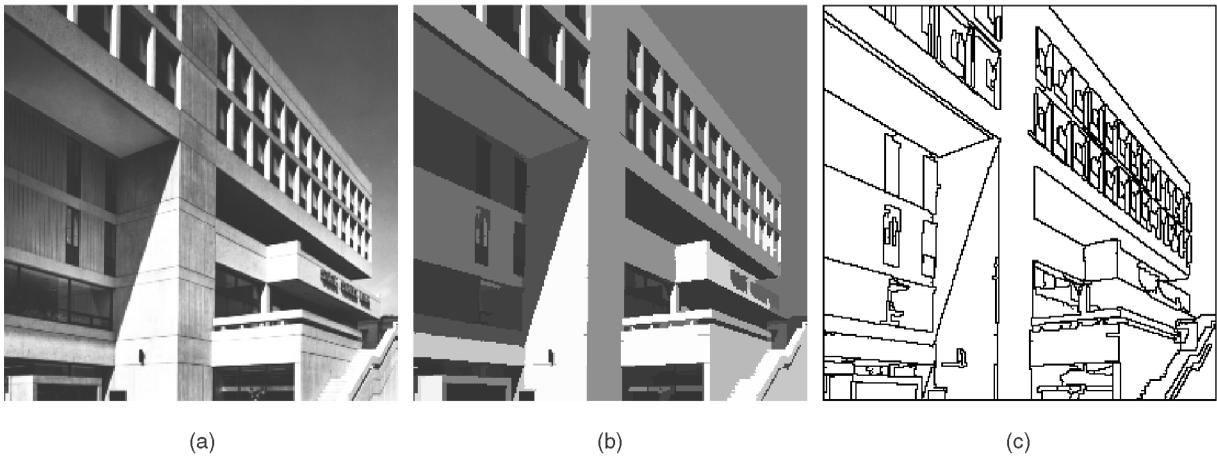


Fig. 6. *MIT* image. (a) Original. (b) Segmented $(h_s, h_r, M) = (8, 7, 20)$. (c) Region boundaries.

introduce artifacts. Recall that the direction of the mean shift is dictated solely by the data.

The mean shift procedure-based image segmentation is a straightforward extension of the discontinuity preserving smoothing algorithm. Each pixel is associated with a *significant* mode of the joint domain density located in its neighborhood, after nearby modes were pruned as in the generic feature space analysis technique (Section 3).

4.2.1 Mean Shift Segmentation

Let \mathbf{x}_i and $\mathbf{z}_i, i = 1, \dots, n$, be the d -dimensional input and filtered image pixels in the joint spatial-range domain and L_i the label of the i th pixel in the segmented image.

1. Run the mean shift filtering procedure for the image and store *all* the information about the d -dimensional convergence point in \mathbf{z}_i , i.e., $\mathbf{z}_i = \mathbf{y}_{i,c}$.
2. Delineate in the joint domain the clusters $\{\mathbf{C}_p\}_{p=1\dots m}$ by grouping together *all* \mathbf{z}_i which are closer than h_s in the spatial domain and h_r in the range domain, i.e., concatenate the basins of attraction of the corresponding convergence points.
3. For each $i = 1, \dots, n$, assign $L_i = \{p \mid \mathbf{z}_i \in \mathbf{C}_p\}$.
4. Optional: Eliminate spatial regions containing less than M pixels.

The cluster delineation step can be refined according to a priori information and, thus, physics-based segmentation algorithms, e.g., [2], [35], can be incorporated. Since this process is performed on region adjacency graphs, hierarchical techniques like [36] can provide significant speed-up. The effect of the cluster delineation step is shown in Fig. 4d. Note the fusion into larger homogeneous regions of the result of filtering shown in Fig. 4c. The segmentation step does not add a significant overhead to the filtering process.

The region representation used by the mean shift segmentation is similar to the blob representation employed in [64]. However, while the blob has a parametric description (multivariate Gaussians in both spatial and color domain), the partition generated by the mean shift is characterized by a nonparametric model. An image region is defined by all the pixels associated with the same mode in the joint domain.

In [43], a nonparametric clustering method is described in which, after kernel density estimation with a small bandwidth, the clusters are delineated through concatenation of the detected modes' neighborhoods. The merging process is based on two intuitive measures capturing the variations in the local density. Being a hierarchical clustering technique, the method is computationally expensive; it takes several minutes in MATLAB to analyze a 2,000 pixel subsample of the feature space. The method is not recommended to be used in the joint domain since the measures employed in the merging process become ineffective. Comparing the results for arbitrarily shaped synthetic data [43, Fig. 6] with a similarly challenging example processed with the mean shift method [12, Fig. 1] shows that the use of a hierarchical approach can be successfully avoided in the nonparametric clustering paradigm.

All the segmentation experiments were performed using uniform kernels. The improvement due to joint space analysis can be seen in Fig. 6 where the 256×256 gray-level image *MIT* was processed with $(h_s, h_r, M) = (8, 7, 20)$. A number of 225 homogeneous regions were identified in fractions of a second, most of them delineating semantically meaningful regions like walls, sky, steps, inscription on the building, etc. Compare the results with the segmentation obtained by one-dimensional clustering of the gray-level values in [11, Fig. 4] or by using a Gibbs random fields-based approach [40, Fig. 7].

The joint domain segmentation of the color 256×256 *room* image presented in Fig. 7 is also satisfactory. Compare this result with the segmentation presented in [38, Figs. 3e and 5c] obtained by recursive thresholding. In both these examples, one can notice that regions in which a small gradient of illumination exists (like the sky in the *MIT* or the carpet in the *room* image) were delineated as a single region. Thus, the joint domain mean shift-based segmentation succeeds in overcoming the inherent limitations of methods based only on gray-level or color clustering which typically oversegment small gradient regions.

The segmentation with $(h_s, h_r, M) = (16, 7, 40)$ of the 512×512 color image *lake* is shown in Fig. 8. Compare this

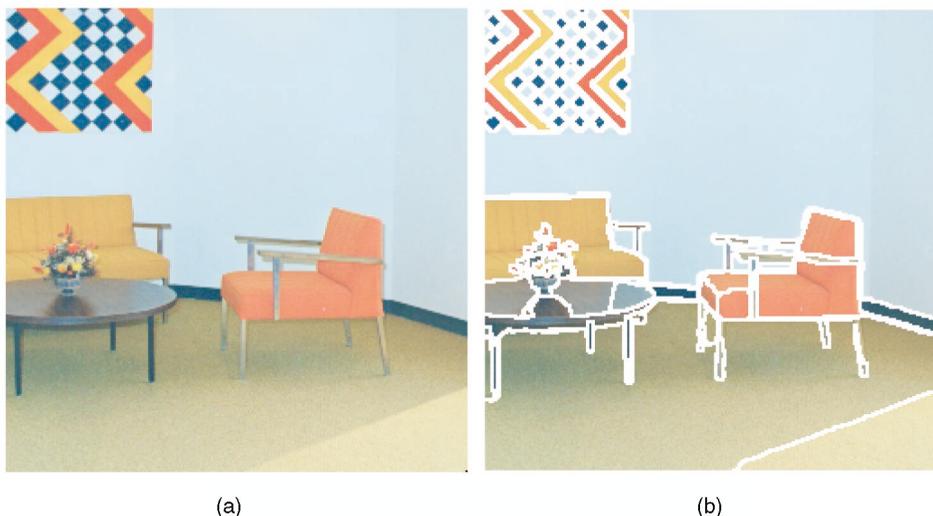


Fig. 7. *Room* image. (a) Original. (b) Region boundaries delineated with $(h_s, h_r, M) = (8, 5, 20)$, drawn over the input.

result with that of the multiscale approach in [57, Fig. 11]. Finally, one can compare the contours of the color image $(h_s, h_r, M) = (16, 19, 40)$ *hand* presented in Fig. 9 with those from [66, Fig. 15], obtained through a complex global optimization, and from [41, Fig. 4a], obtained with geodesic active contours.

The segmentation is not very sensitive to the choice of the resolution parameters h_s and h_r . Note that all 256×256 images used the same $h_s = 8$, corresponding to a 17×17 spatial window, while all 512×512 images used $h_s = 16$ corresponding to a 31×31 window. The range parameter h_r and the smallest significant feature size M control the number of regions in the segmented image. The more an image deviates from the assumed piecewise constant model, larger values have to be used for h_r and M to discard the effect of small local variations in the feature space. For example, the heavily textured background in the *hand* image is compensated by using $h_r = 19$ and $M = 40$, values which are much larger than those used for the *room* image

$(h_r = 5, M = 20)$ since the latter better obeys the model. As with any low-level vision algorithm, the quality of the segmentation output can be assessed only in the context of the whole vision task and, thus, the resolution parameters should be chosen according to that criterion. An important advantage of mean shift-based segmentation is its modularity which makes the control of segmentation output very simple.

Other segmentation examples in which the original image has the region boundaries superposed are shown in Fig. 10 and in which the original and labeled images are compared in Fig. 11.

As a potential application of the segmentation, we return to the *cameraman* image. Fig. 12a shows the reconstructed image after the regions corresponding to the sky and grass were manually replaced with white. The mean shift segmentation has been applied with $(h_s, h_r, M) = (8, 4, 10)$. Observe the preservation of the details which suggests that the algorithm can also be used for image editing, as shown in Fig. 12b.

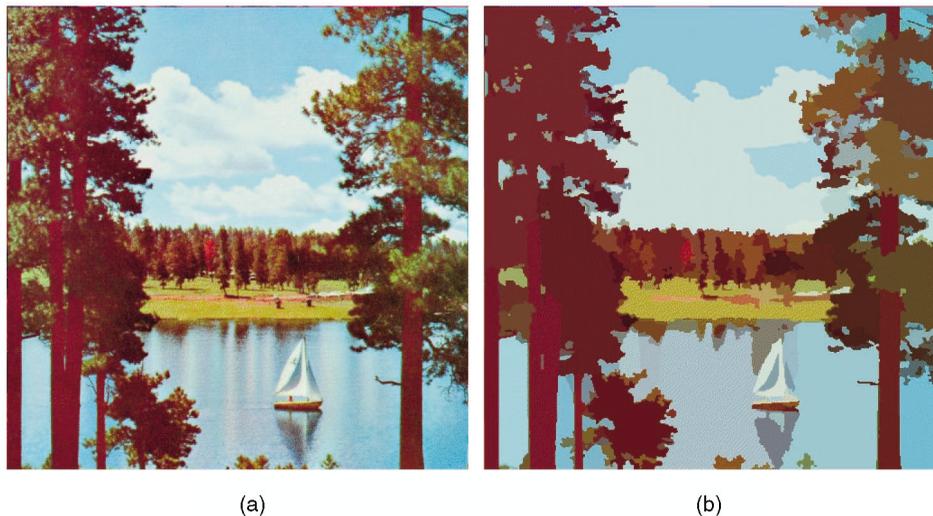


Fig. 8. *Lake* image. (a) Original. (b) Segmented with $(h_s, h_r, M) = (16, 7, 40)$.

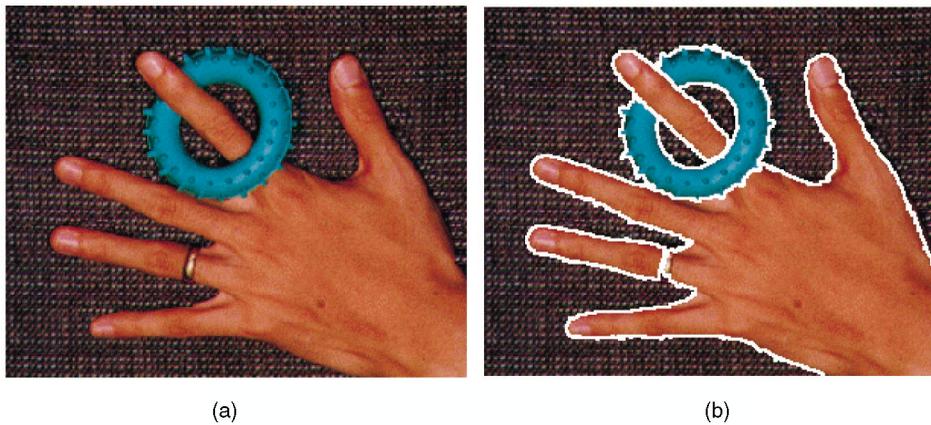


Fig. 9. *Hand* image. (a) Original. (b) Region boundaries delineated with $(h_s, h_r, M) = (16, 19, 40)$ drawn over the input.

The code for the discontinuity preserving smoothing and image segmentation algorithms integrated into a single system with graphical interface is available at <http://www.caip.rutgers.edu/riul/research/code.html>.

5 DISCUSSION

The mean shift-based feature space analysis technique introduced in this paper is a general tool which is not restricted to the two applications discussed here. Since the quality of the output is controlled only by the kernel bandwidth, i.e., the resolution of the analysis, the technique should be also easily integrable into complex vision systems where the control is relinquished to a closed loop process. Additional insights on the bandwidth selection can be obtained by testing the stability of the mean shift direction across the different bandwidths, as investigated in [57] in

the case of the force field. The nonparametric toolbox developed in this paper is suitable for a large variety of computer vision tasks where parametric models are less adequate, for example, modeling the background in visual surveillance [18].

The complete solution toward autonomous image segmentation is to combine a bandwidth selection technique (like the ones discussed in Section 3.1) with top-down task-related high-level information. In this case, each mean shift process is associated with a kernel best suited to the local structure of the joint domain. Several interesting theoretical issues have to be addressed, though, before the benefits of such a data driven approach can be fully exploited. We are currently investigating these issues.

The ability of the mean shift procedure to be attracted by the modes (local maxima) of an underlying density function, can be exploited in an optimization framework. Cheng [7]



Fig. 10. *Landscape* images. All the region boundaries were delineated with $(h_s, h_r, M) = (8, 7, 100)$ and are drawn over the original image.

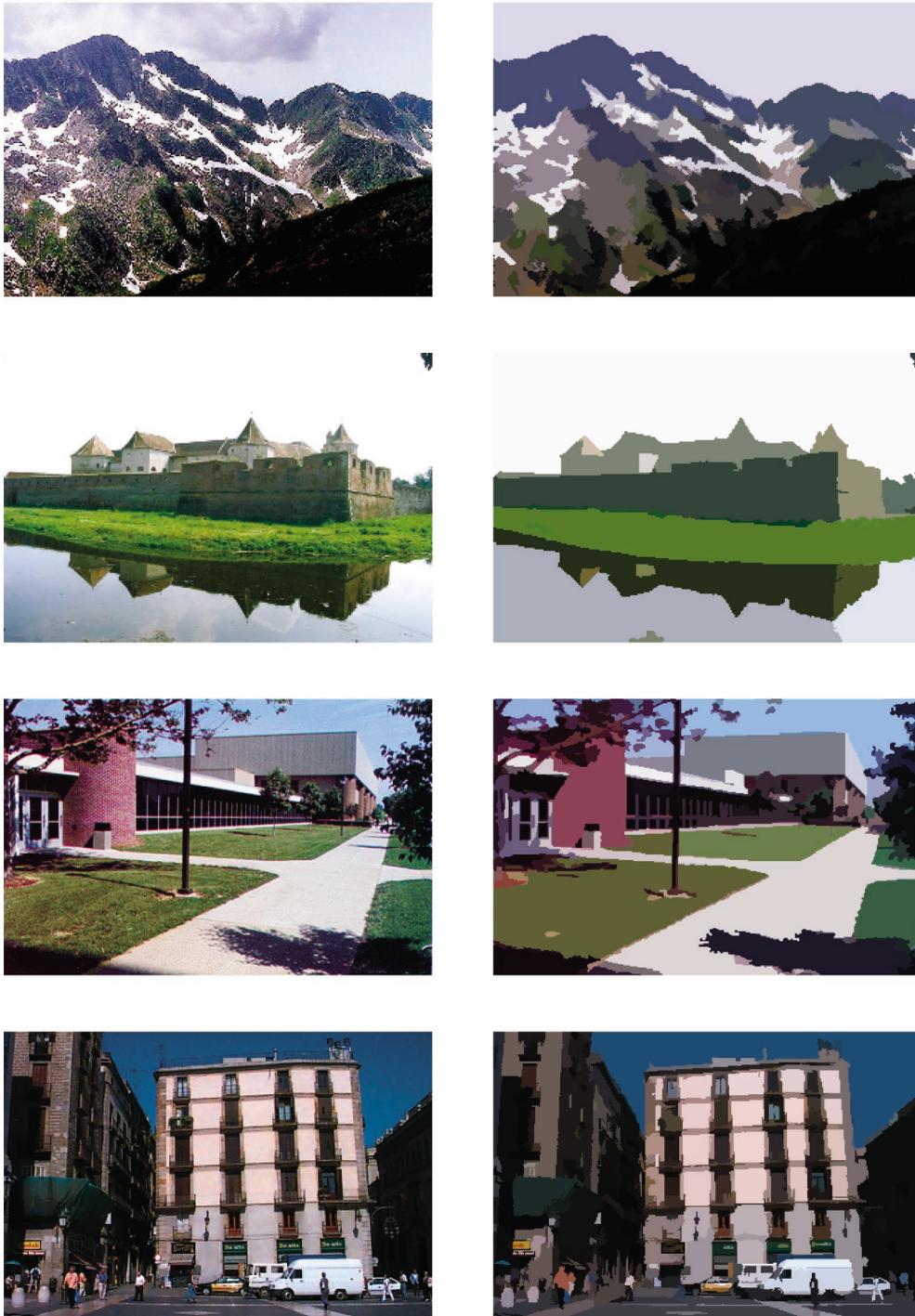


Fig. 11. Some other segmentation examples with $(h_s, h_r, M) = (8, 7, 20)$. Left: original. Right: segmented.

already discusses a simple example. However, by introducing adequate objective functions, the optimization problem can acquire physical meaning in the context of a computer vision task. For example, in [14], by defining the distance between the distributions of the model and a candidate of the target, nonrigid objects were tracked in an image sequence under severe distortions. The distance was defined at every pixel in the region of interest of the new frame and the mean shift procedure was used to find the mode of this measure nearest to the previous location of the target.

The above-mentioned tracking algorithm can be regarded as an example of computer vision techniques which are based on *in situ* optimization. Under this paradigm, the solution is obtained by using the input domain to define the optimization problem. The *in situ* optimization is a very powerful method. In [23] and [58], each input data point was associated with a local field (voting kernel) to produce a more dense structure from where the sought information (salient features, the hyperplane representing the fundamental matrix) can be reliably extracted.

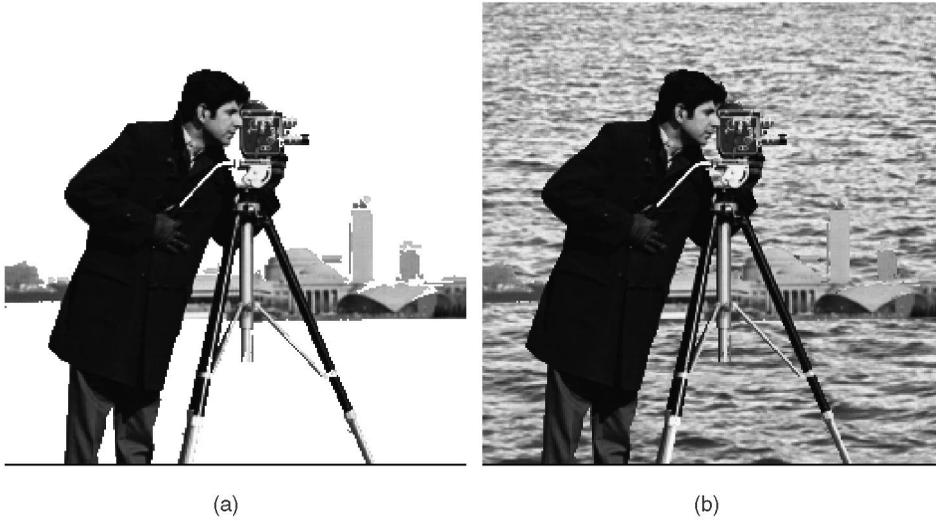


Fig. 12. *Cameraman* image. (a) Segmentation with $(h_s, h_r, M) = (8, 4, 10)$ and reconstruction after the elimination of regions representing sky and grass. (b) Supervised texture insertion.

The mean shift procedure is not computationally expensive. Careful C++ implementation of the tracking algorithm allowed real time (30 frames/second) processing of the video stream. While it is not clear if the segmentation algorithm described in this paper can be made so fast, given the quality of the region boundaries it provides, it can be used to support edge detection without significant overhead in time.

Kernel density estimation, in particular, and nonparametric techniques, in general, do not scale well with the dimension of the space. This is mostly due to the empty space phenomenon [20, p. 70], [54, p. 93] by which most of the mass in a high-dimensional space is concentrated in a small region of the space. Thus, whenever the feature space has more than (say) six dimensions, the analysis should be approached carefully. Employing projection pursuit, in which the density is analyzed along lower dimensional cuts, e.g., [27], is a possibility.

To conclude, the mean shift procedure is a valuable computational module whose versatility can make it an important component of any computer vision toolbox.

APPENDIX

Proof of Theorem 1. *If the kernel K has a convex and monotonically decreasing profile, the sequences $\{\mathbf{y}_j\}_{j=1,2,\dots}$ and $\{\hat{f}_{h,K}(j)\}_{j=1,2,\dots}$ converge, and $\{\hat{f}_{h,K}(j)\}_{j=1,2,\dots}$ is monotonically increasing.*

Since n is finite, the sequence $\hat{f}_{h,K}$ (21) is bounded, therefore, it is sufficient to show that $\hat{f}_{h,K}$ is strictly monotonic increasing, i.e., if $\mathbf{y}_j \neq \mathbf{y}_{j+1}$, then

$$\hat{f}_{h,K}(j) < \hat{f}_{h,K}(j+1),$$

for $j = 1, 2, \dots$. Without loss of generality, it can be assumed that $\mathbf{y}_j = \mathbf{0}$ and, thus, from (16) and (21)

$$\begin{aligned} \hat{f}_{h,K}(j+1) - \hat{f}_{h,K}(j) = \\ \frac{c_{k,d}}{nh^d} \sum_{i=1}^n \left[k\left(\left\|\frac{\mathbf{y}_{j+1} - \mathbf{x}_i}{h}\right\|^2\right) - k\left(\left\|\frac{\mathbf{x}_i}{h}\right\|^2\right) \right]. \end{aligned} \quad (\text{A.1})$$

The convexity of the profile $k(x)$ implies that

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

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

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

Now, using (A.1) and (A.3), we obtain

$$\begin{aligned} \hat{f}_{h,K}(j+1) - \hat{f}_{h,K}(j) \\ \geq \frac{c_{k,d}}{nh^{d+2}} \sum_{i=1}^n g\left(\left\|\frac{\mathbf{x}_i}{h}\right\|^2\right) \left[\|\mathbf{x}_i\|^2 - \|\mathbf{y}_{j+1} - \mathbf{x}_i\|^2 \right] \\ = \frac{c_{k,d}}{nh^{d+2}} \sum_{i=1}^n g\left(\left\|\frac{\mathbf{x}_i}{h}\right\|^2\right) \left[2\mathbf{y}_{j+1}^\top \mathbf{x}_i - \|\mathbf{y}_{j+1}\|^2 \right] \\ = \frac{c_{k,d}}{nh^{d+2}} \left[2\mathbf{y}_{j+1}^\top \sum_{i=1}^n \mathbf{x}_i g\left(\left\|\frac{\mathbf{x}_i}{h}\right\|^2\right) - \|\mathbf{y}_{j+1}\|^2 \sum_{i=1}^n g\left(\left\|\frac{\mathbf{x}_i}{h}\right\|^2\right) \right] \end{aligned} \quad (\text{A.4})$$

and, recalling (20), yields

$$\hat{f}_{h,K}(j+1) - \hat{f}_{h,K}(j) \geq \frac{c_{k,d}}{nh^{d+2}} \|\mathbf{y}_{j+1}\|^2 \sum_{i=1}^n g\left(\left\|\frac{\mathbf{x}_i}{h}\right\|^2\right). \quad (\text{A.5})$$

The profile $k(x)$ being monotonically decreasing for all $x \geq 0$, the sum $\sum_{i=1}^n g\left(\left\|\frac{\mathbf{x}_i}{h}\right\|^2\right)$ is strictly positive. Thus, as long as $\mathbf{y}_{j+1} \neq \mathbf{0}$, the right term of (A.5) is strictly positive, i.e., $\hat{f}_{h,K}(j+1) > \hat{f}_{h,K}(j)$. Consequently, the sequence $\{\hat{f}_{h,K}(j)\}_{j=1,2,\dots}$ is convergent.

To prove the convergence of the sequence $\{\mathbf{y}_j\}_{j=1,2,\dots}$, (A.5) is rewritten for an arbitrary kernel location $\mathbf{y}_j \neq \mathbf{0}$. After some algebra, we have

$$\hat{f}_{h,K}(j+1) - \hat{f}_{h,K}(j) \geq \frac{c_{k,d}}{nh^{d+2}} \|\mathbf{y}_{j+1} - \mathbf{y}_j\|^2 \sum_{i=1}^n g\left(\left\|\frac{\mathbf{y}_j - \mathbf{x}_i}{h}\right\|^2\right). \quad (\text{A.6})$$

Now, summing the two terms of (A.6) for indices $j, j+1, \dots, j+m-1$, it results that

$$\begin{aligned}
& \hat{f}_{h,K}(j+m) - \hat{f}_{h,K}(j) \\
& \geq \frac{C_{k,d}}{nh^{d+2}} \|\mathbf{y}_{j+m} - \mathbf{y}_{j+m-1}\|^2 \sum_{i=1}^n g\left(\left\|\frac{\mathbf{y}_{j+m-1} - \mathbf{x}_i}{h}\right\|^2\right) + \dots \\
& \quad + \frac{C_{k,d}}{nh^{d+2}} \|\mathbf{y}_{j+1} - \mathbf{y}_j\|^2 \sum_{i=1}^n g\left(\left\|\frac{\mathbf{y}_j - \mathbf{x}_i}{h}\right\|^2\right) \\
& \geq \frac{C_{k,d}}{nh^{d+2}} \left[\|\mathbf{y}_{j+m} - \mathbf{y}_{j+m-1}\|^2 + \dots + \|\mathbf{y}_{j+1} - \mathbf{y}_j\|^2 \right] M \\
& \geq \frac{C_{k,d}}{nh^{d+2}} \|\mathbf{y}_{j+m} - \mathbf{y}_j\|^2 M,
\end{aligned} \tag{A.7}$$

where M represents the minimum (always strictly positive) of the sum $\sum_{i=1}^n g\left(\left\|\frac{\mathbf{y}_j - \mathbf{x}_i}{h}\right\|^2\right)$ for all $\{\mathbf{y}_j\}_{j=1,2,\dots}$.

Since $\{\hat{f}_{h,K}(j)\}_{j=1,2,\dots}$ is convergent, it is also a Cauchy sequence. This property in conjunction with (A.7) implies that $\{\mathbf{y}_j\}_{j=1,2,\dots}$ is a Cauchy sequence, hence, it is convergent in the Euclidean space. \square

Proof of Theorem 2. *The cosine of the angle between two consecutive mean shift vectors is strictly positive when a normal kernel is employed.*

We can assume, without loss of generality that $\mathbf{y}_j = 0$ and $\mathbf{y}_{j+1} \neq \mathbf{y}_{j+2} \neq \mathbf{0}$ since, otherwise, convergence has already been achieved. Therefore, the mean shift vector $\mathbf{m}_{h,N}(\mathbf{0})$ is

$$\mathbf{m}_{h,N}(\mathbf{0}) = \mathbf{y}_{j+1} = \frac{\sum_{i=1}^n \mathbf{x}_i \exp\left(-\left\|\frac{\mathbf{x}_i}{h}\right\|^2\right)}{\sum_{i=1}^n \exp\left(-\left\|\frac{\mathbf{x}_i}{h}\right\|^2\right)}. \tag{B.1}$$

We will show first that, when the weights are given by a normal kernel centered at \mathbf{y}_{j+1} , the weighted sum of the projections of $(\mathbf{y}_{j+1} - \mathbf{x}_i)$ onto \mathbf{y}_{j+1} is strictly negative, i.e.,

$$\sum_{i=1}^n \left(\|\mathbf{y}_{j+1}\|^2 - \mathbf{y}_{j+1}^\top \mathbf{x}_i \right) \exp\left(-\left\|\frac{\mathbf{y}_{j+1} - \mathbf{x}_i}{h}\right\|^2\right) < 0. \tag{B.2}$$

The space R^d can be decomposed into the following three domains:

$$\begin{aligned}
D_1 &= \left\{ \mathbf{x} \in R^d \mid \mathbf{y}_{j+1}^\top \mathbf{x} \leq \frac{1}{2} \|\mathbf{y}_{j+1}\|^2 \right\} \\
D_2 &= \left\{ \mathbf{x} \in R^d \mid \frac{1}{2} \|\mathbf{y}_{j+1}\|^2 < \mathbf{y}_{j+1}^\top \mathbf{x} \leq \|\mathbf{y}_{j+1}\|^2 \right\} \\
D_3 &= \left\{ \mathbf{x} \in R^d \mid \|\mathbf{y}_{j+1}\|^2 < \mathbf{y}_{j+1}^\top \mathbf{x} \right\}
\end{aligned} \tag{B.3}$$

and after some simple manipulations from (B.1), we can derive the equality

$$\begin{aligned}
& \sum_{\mathbf{x}_i \in D_2} \left(\|\mathbf{y}_{j+1}\|^2 - \mathbf{y}_{j+1}^\top \mathbf{x}_i \right) \exp\left(-\left\|\frac{\mathbf{x}_i}{h}\right\|^2\right) \\
&= \sum_{\mathbf{x}_i \in D_1 \cup D_3} \left(\mathbf{y}_{j+1}^\top \mathbf{x}_i - \|\mathbf{y}_{j+1}\|^2 \right) \exp\left(-\left\|\frac{\mathbf{x}_i}{h}\right\|^2\right).
\end{aligned} \tag{B.4}$$

In addition, for $\mathbf{x} \in D_2$, we have $\|\mathbf{y}_{j+1}\|^2 - \mathbf{y}_{j+1}^\top \mathbf{x} \geq 0$, which implies

$$\|\mathbf{y}_{j+1} - \mathbf{x}_i\|^2 = \|\mathbf{y}_{j+1}\|^2 + \|\mathbf{x}_i\|^2 - 2\mathbf{y}_{j+1}^\top \mathbf{x}_i \geq \|\mathbf{x}_i\|^2 - \|\mathbf{y}_{j+1}\|^2 \tag{B.5}$$

from where

$$\begin{aligned}
& \sum_{\mathbf{x}_i \in D_2} \left(\|\mathbf{y}_{j+1}\|^2 - \mathbf{y}_{j+1}^\top \mathbf{x}_i \right) \exp\left(-\left\|\frac{\mathbf{y}_{j+1} - \mathbf{x}_i}{h}\right\|^2\right) \\
& \leq \exp\left(\left\|\frac{\mathbf{y}_{j+1}}{h}\right\|^2\right) \sum_{\mathbf{x}_i \in D_2} \left(\|\mathbf{y}_{j+1}\|^2 - \mathbf{y}_{j+1}^\top \mathbf{x}_i \right) \exp\left(-\left\|\frac{\mathbf{x}_i}{h}\right\|^2\right).
\end{aligned} \tag{B.6}$$

Now, introducing (B.4) in (B.6), we have

$$\begin{aligned}
& \sum_{\mathbf{x}_i \in D_2} \left(\|\mathbf{y}_{j+1}\|^2 - \mathbf{y}_{j+1}^\top \mathbf{x}_i \right) \exp\left(-\left\|\frac{\mathbf{y}_{j+1} - \mathbf{x}_i}{h}\right\|^2\right) \\
& \leq \exp\left(\left\|\frac{\mathbf{y}_{j+1}}{h}\right\|^2\right) \sum_{\mathbf{x}_i \in D_1 \cup D_3} \left(\mathbf{y}_{j+1}^\top \mathbf{x}_i - \|\mathbf{y}_{j+1}\|^2 \right) \exp\left(-\left\|\frac{\mathbf{x}_i}{h}\right\|^2\right)
\end{aligned} \tag{B.7}$$

and, by adding to both sides of (B.7), the quantity

$$\sum_{\mathbf{x}_i \in D_1 \cup D_3} \left(\|\mathbf{y}_{j+1}\|^2 - \mathbf{y}_{j+1}^\top \mathbf{x}_i \right) \exp\left(-\left\|\frac{\mathbf{y}_{j+1} - \mathbf{x}_i}{h}\right\|^2\right),$$

after some algebra, it results that

$$\begin{aligned}
& \sum_{i=1}^n \left(\|\mathbf{y}_{j+1}\|^2 - \mathbf{y}_{j+1}^\top \mathbf{x}_i \right) \exp\left(-\left\|\frac{\mathbf{y}_{j+1} - \mathbf{x}_i}{h}\right\|^2\right) \\
& \leq \exp\left(\left\|\frac{\mathbf{y}_{j+1}}{h}\right\|^2\right) \sum_{\mathbf{x}_i \in D_1 \cup D_3} \left(\|\mathbf{y}_{j+1}\|^2 - \mathbf{y}_{j+1}^\top \mathbf{x}_i \right) \exp\left(-\left\|\frac{\mathbf{x}_i}{h}\right\|^2\right) \\
& \quad \times \left\{ \exp\left[-\frac{2}{h^2} \left(\|\mathbf{y}_{j+1}\|^2 - \mathbf{y}_{j+1}^\top \mathbf{x}_i \right) \right] - 1 \right\}.
\end{aligned} \tag{B.8}$$

The right side of (B.8) is negative because

$$\left(\|\mathbf{y}_{j+1}\|^2 - \mathbf{y}_{j+1}^\top \mathbf{x}_i \right)$$

and the last product term has opposite signs in both the D_1 and D_3 domains. Therefore, the left side of (B.8) is also negative, which proves (B.2).

We can use now (B.2) to write

$$\|\mathbf{y}_{j+1}\|^2 < \mathbf{y}_{j+1}^\top \frac{\sum_{i=1}^n \mathbf{x}_i \exp\left(-\left\|\frac{\mathbf{y}_{j+1} - \mathbf{x}_i}{h}\right\|^2\right)}{\sum_{i=1}^n \exp\left(-\left\|\frac{\mathbf{y}_{j+1} - \mathbf{x}_i}{h}\right\|^2\right)} = \mathbf{y}_{j+1}^\top \mathbf{y}_{j+2}, \tag{B.9}$$

from where

$$\frac{\mathbf{y}_{j+1}^\top (\mathbf{y}_{j+2} - \mathbf{y}_{j+1})}{\|\mathbf{y}_{j+1}\| \|\mathbf{y}_{j+2} - \mathbf{y}_{j+1}\|} > 0 \tag{B.10}$$

or by taking into account (24)

$$\frac{\mathbf{m}_{h,N}(\mathbf{y}_j)^\top \mathbf{m}_{h,N}(\mathbf{y}_{j+1})}{\|\mathbf{m}_{h,N}(\mathbf{y}_j)\| \|\mathbf{m}_{h,N}(\mathbf{y}_{j+1})\|} > 0.$$

\square

ACKNOWLEDGMENTS

The support of the US National Science Foundation under grants IRI 95-30546 and IRI 99-87695 is gratefully

acknowledged. Preliminary versions for some parts of the material were presented in [13] and [14]. The authors would like to thank John Kent from the University of Leeds and David Tyler of Rutgers for discussions about the relation between the mean shift procedure and M-estimators.

REFERENCES

- [1] G. Aggarwal, S. Ghosal, and P. Dubey, "Efficient Query Modification for Image Retrieval," *Proc. 2000 IEEE Conf. Computer Vision and Pattern Recognition*, vol. II, pp. 255-261, June 2000.
- [2] R. Bajcsy, S.W. Lee, and A. Leonardis, "Detection of Diffuse and Specular Interface Reflections and Inter-Reflections by Color Image Segmentation," *Int'l J. Computer Vision*, vol. 17, pp. 241-272, 1996.
- [3] D. Barash, "Bilateral Filtering and Anisotropic Diffusion: Towards a Unified Viewpoint," *IEEE Trans. Pattern Analysis and Pattern Recognition*, to appear.
- [4] D.P. Bertsekas, *Nonlinear Programming*. Athena Scientific, 1995.
- [5] M.J. Black, G. Sapiro, D.H. Marimont, and D. Heeger, "Robust Anisotropic Diffusion," *IEEE Trans. Image Processing*, vol. 7, pp. 421-432, 1998.
- [6] G.R. Bradski, "Computer Vision Face Tracking as a Component of a Perceptual User Interface," *Proc. IEEE Workshop Applications of Computer Vision*, pp. 214-219, Oct. 1998.
- [7] Y. Cheng, "Mean Shift, Mode Seeking, and Clustering," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 17, no. 8, pp. 790-799, Aug. 1995.
- [8] E. Choi and P. Hall, "Data Sharpening as a Prelude to Density Estimation," *Biometrika*, vol. 86, pp. 941-947, 1999.
- [9] C.K. Chu, I.K. Glad, F. Godtlielsen, and J.S. Maron, "Edge-Preserving Smoothers for Image Processing," *J. Am. Statistical Assoc.*, vol. 93, pp. 526-541, 1998.
- [10] D. Comaniciu, "Nonparametric Robust Methods for Computer Vision," PhD thesis, Dept. of Electrical and Computer Eng., Rutgers Univ., 1999. Available at <http://www.caip.rutgers.edu/riul/research/theses.html>.
- [11] D. Comaniciu and P. Meer, "Robust Analysis of Feature Spaces: Color Image Segmentation," *Proc. 1997 IEEE Conf. Computer Vision and Pattern Recognition*, pp. 750-755, June 1997.
- [12] D. Comaniciu and P. Meer, "Distribution Free Decomposition of Multivariate Data," *Pattern Analysis and Applications*, vol. 2, pp. 22-30, 1999.
- [13] D. Comaniciu and P. Meer, "Mean Shift Analysis and Applications," *Proc. Seventh Int'l Conf. Computer Vision*, pp. 1197-1203, Sept. 1999.
- [14] D. Comaniciu, V. Ramesh, and P. Meer, "Real-Time Tracking of Non-Rigid Objects Using Mean Shift," *Proc. 2000 IEEE Conf. Computer Vision and Pattern Recognition*, vol. II, pp. 142-149, June 2000.
- [15] D. Comaniciu, V. Ramesh, and P. Meer, "The Variable Bandwidth Mean Shift and Data-Driven Scale Selection," *Proc. Eighth Int'l Conf. Computer Vision*, vol. I, pp. 438-445, July 2001.
- [16] C. Connolly, "The Relationship between Colour Metrics and the Appearance of Three-Dimensional Coloured Objects," *Color Research and Applications*, vol. 21, pp. 331-337, 1996.
- [17] R.O. Duda and P.E. Hart, *Pattern Classification and Scene Analysis*. Wiley, 1973.
- [18] A. Elgammal, D. Harwood, and L. Davis, "Non-Parametric Model for Background Subtraction," *Proc. Sixth European Conf. Computer Vision*, vol. II, pp. 751-767, June 2000.
- [19] B. Fischl and E.L. Schwartz, "Adaptive Nonlocal Filtering: A Fast Alternative to Anisotropic Diffusion for Image Enhancement," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 21, no. 1, pp. 42-48, Jan. 1999.
- [20] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, second ed. Academic Press, 1990.
- [21] K. Fukunaga and L.D. Hostetler, "The Estimation of the Gradient of a Density Function, with Applications in Pattern Recognition," *IEEE Trans. Information Theory*, vol. 21, pp. 32-40, 1975.
- [22] J. Guo, J. Kim, and C. Kuo, "Fast and Accurate Moving Object Extraction Technique for MPEG-4 Object Based Video Coding," *Proc. SPIE Visual Comm. and Image Processing*, vol. 3653, pp. 1210-1221, 1999.
- [23] G. Guy and G. Medioni, "Inference of Surfaces, 3D Curves, and Junctions from Sparse, Noisy, 3D Data," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 19, pp. 1265-1277, 1997.
- [24] W. Härdle, *Applied Nonparametric Regression*. Cambridge Univ. Press, 1991.
- [25] M. Herbin, N. Bonnet, and P. Vautrot, "A Clustering Method Based on the Estimation of the Probability Density Function and on the Skeleton by Influence Zones," *Pattern Recognition Letters*, vol. 17, pp. 1141-1150, 1996.
- [26] P.J. Huber, *Robust Statistical Procedures*, second ed. SIAM, 1996.
- [27] J.N. Hwang and S.R. Lay, and A. Lippman, "Nonparametric Multivariate Density Estimation: A Comparative Study," *IEEE Trans. Signal Processing*, vol. 42, pp. 2795-2810, 1994.
- [28] A.K. Jain and R.C. Dubes, *Algorithms for Clustering Data*. Prentice Hall, 1988.
- [29] A.K. Jain, R.P.W. Duin, and J. Mao, "Statistical Pattern Recognition: A Review," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 22, no. 1, pp. 4-37, Jan. 2000.
- [30] L. Kauffman and P. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*. J. Wiley & Sons, 1990.
- [31] D.T. Kuan, A.A. Sawchuk, T.C. Strand, and P. Chavel, "Adaptive Noise Smoothing Filter for Images with Signal Dependent Noise," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 7, no. 2, pp. 165-177, Mar. 1985.
- [32] G. Li, "Robust Regression," *Exploring Data Tables, Trends, and Shapes*, D.C. Hoaglin, F. Mosteller, and J.W. Tukey, eds., pp. 281-343. Wiley, 1985.
- [33] L. Liu and S. Sclaroff, "Deformable Shape Detection and Description via Model-Based Region Grouping," *Proc. 1999 IEEE Conf. Computer Vision and Pattern Recognition*, vol. II, pp. 21-27, June 1999.
- [34] W.Y. Ma and B.S. Manjunath, "Edge flow: A Framework of Boundary Detection and Image Segmentation," *IEEE Trans. Image Processing*, vol. 9, pp. 1375-1388, 2000.
- [35] B.A. Maxwell and S.S. Shafer, "Segmentation and Interpretation of Multicolored Objects with Highlights," *Computer Vision and Image Understanding*, vol. 77, pp. 1-24, 2000.
- [36] A. Montanvert, P. Meer, and A. Rosenfeld, "Hierarchical Image Analysis Using Irregular Tessellation," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 13, no. 4, pp. 307-316, Apr. 1991.
- [37] J. Monteil and A. Beghdadi, "A New Interpretation and Improvement of Nonlinear Anisotropic Diffusion for Image Enhancement," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 21, no. 9, pp. 940-946, Sept. 1999.
- [38] Y. Ohta, T. Kanade, and T. Sakai, "Color Information for Region Segmentation," *Computer Graphics and Image Processing*, vol. 13, pp. 222-241, 1980.
- [39] J. Pan, F. McInnes, and M. Jack, "Fast Clustering Algorithms for Vector Quantization," *Pattern Recognition*, vol. 29, pp. 511-518, 1996.
- [40] T.N. Pappas, "An Adaptive Clustering Algorithm for Image Segmentation," *IEEE Trans. Signal Processing*, vol. 40, pp. 901-914, 1992.
- [41] N. Paragios and R. Deriche, "Geodesic Active Contours for Supervised Texture Segmentation," *Proc. 1999 IEEE Conf. Computer Vision and Pattern Recognition*, vol. II, pp. 422-427, June 1999.
- [42] B. Park and J. Marron, "Comparison of Data-Driven Bandwidth Selectors," *J. Am. Statistical Assoc.*, vol. 85, pp. 66-72, 1990.
- [43] E.J. Pauwels and G. Frederix, "Finding Salient Regions in Images," *Computer Vision and Image Understanding*, vol. 75, pp. 73-85, 1999.
- [44] P. Perona and J. Malik, "Scale-Space and Edge Detection Using Anisotropic Diffusion," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 12, no. 7, pp. 629-639, July 1990.
- [45] K. Popat and R.W. Picard, "Cluster-Based Probability Model and Its Application to Image and Texture Processing," *IEEE Trans. Image Processing*, vol. 6, pp. 268-284, 1997.
- [46] W.K. Pratt, *Digital Image Processing*, second ed. Wiley, 1991.
- [47] D. Ridder, J. Kittler, O. Lemmers, and R. Duin, "The Adaptive Subspace Map for Texture Segmentation," *Proc. 2000 Int'l Conf. Pattern Recognition*, pp. 216-220, Sept. 2000.
- [48] T. Risse, "Hough Transform for Line Recognition: Complexity of Evidence Accumulation and Cluster Detection," *Computer Vision Graphics and Image Processing*, vol. 46, pp. 327-345, 1989.
- [49] S.J. Roberts, "Parametric and Non-Parametric Unsupervised Cluster Analysis," *Pattern Recognition*, vol. 30, pp. 261-272, 1997.
- [50] P. Saint-Marc and J.S. Chen, G. Medioni, "Adaptive Smoothing: A General Tool for Early Vision," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 13, no. 6, pp. 514-529, June 1991.
- [51] D.W. Scott, *Multivariate Density Estimation*. Wiley, 1992.
- [52] R. Sedgewick, *Algorithms in C++*. Addison-Wesley, 1992.

- [53] S. Sheather, M. Jones, "A Reliable Data-Based Bandwidth Selection Method for Kernel Density Estimation," *J. Royal Statistics Soc. B*, vol. 53, pp. 683-690, 1991.
- [54] B. W. Silverman, *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, 1986.
- [55] J. Simonoff, *Smoothing Methods in Statistics*. Springer-Verlag, 1996.
- [56] Special Issue on Partial Differential Equations and Geometry-Driven Diffusion in Image Processing and Analysis, *IEEE Trans. Image Processing*, vol. 7, Mar. 1998.
- [57] M. Tabb and N. Ahuja, "Multiscale Image Segmentation by Integrated Edge and Region Detection," *IEEE Trans. Image Processing*, vol. 6, pp. 642-655, 1997.
- [58] C.K. Tang, G. Medioni, and M.S. Lee, "Epipolar Geometry Estimation by Tensor Voting in 8D," *Proc. Seventh Int'l Conf. Computer Vision*, vol. I, pp. 502-509, Sept. 1999.
- [59] C. Tomasi and R. Manduchi, "Bilateral Filtering for Gray and Color Images," *Proc. Sixth Int'l Conf. Computer Vision*, pp. 839-846, Jan. 1998.
- [60] A. Touzani and J.G. Postaire, "Clustering by Mode Boundary Detection," *Pattern Recognition Letters*, vol. 9, pp. 1-12, 1989.
- [61] T. Tuytelaars, L. Van Gool, M. Proesmans, and T. Moons, "The Cascaded Hough Transform as an Aid in Aerial Image Interpretation," *Proc. Sixth Int'l Conf. Computer Vision*, pp. 67-72, Jan. 1998.
- [62] M.P. Wand and M. Jones, *Kernel Smoothing*. Chapman and Hall, 1995.
- [63] R. Wilson and M. Spann, "A New Approach to Clustering," *Pattern Recognition*, vol. 23, pp. 1413-1425, 1990.
- [64] C. Wren, A. Azarbayejani, T. Darrell, and A. Pentland, "Pfinder: Real-Time Tracking of the Human Body," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 780-785, July 1997.
- [65] G. Wyszecki and W.S. Stiles, *Color Science: Concepts and Methods, Quantitative Data and Formulae*. second ed. Wiley, 1982.
- [66] S.C. Zhu and A. Yuille, "Region Competition: Unifying Snakes, Region Growing, and Bayes/MDL for Multiband Image Segmentation," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 18, no. 9, pp. 884-900, Sept. 1996.
- [67] X. Zhuang, Y. Huang, K. Palaniappan, and Y. Zhao, "Gaussian Mixture Density Modeling Decomposition, and Applications," *IEEE Trans. Image Processing*, vol. 5, pp. 1293-1302, 1996.



Dorin Comaniciu received the Dipl. Engn. and PhD degrees in electronics from the Polytechnic University of Bucharest in 1988 and 1995 and the PhD degree in electrical engineering from Rutgers University in 1999. From 1988 to 1990, he was with ICE Felix Computers in Bucharest. In 1991, he joined the Department of Electronics and Telecommunications at the Polytechnic University of Bucharest and he held research appointments in Germany and France. From 1996 to 1999, he was with the Center for Advanced Information Processing associated with Rutgers University. Since 1999, he has been a member of the technical staff at Siemens Corporate Research in Princeton, New Jersey. His research interests include robust methods for autonomous computer vision, nonparametric analysis, real-time vision systems, video surveillance, content-based access to visual data, and data compression. He has coauthored numerous papers, conference papers, and book chapters in the area of visual information processing. He received the Best Paper Award at the IEEE Conference Computer Vision and Pattern Recognition 2000. He is a member of the IEEE.



Peter Meer received the Dipl. Engn. degree from the Bucharest Polytechnic Institute, Bucharest, Romania, in 1971, and the DSc degree from the Technion, Israel Institute of Technology, Haifa, Israel, in 1986, both in electrical engineering. From 1971 to 1979, he was with the Computer Research Institute, Cluj, Romania, working on research and development of digital hardware. From 1986 to 1990, he was an assistant research scientist at the Center for Automation

Research, University of Maryland at College Park. In 1991, he joined the Department of Electrical and Computer Engineering, Rutgers University, Piscataway, New Jersey, where he is currently an associate professor. He has held visiting appointments in Japan, Korea, Sweden, Israel and France and was on the organizing committees of numerous International workshops and conferences. He is an associate editor of the *IEEE Transaction on Pattern Analysis and Machine Intelligence*, a member of the editorial board of *Pattern Recognition*, and he was a guest editor of *Computer Vision and Image Understanding* for the April 2000 special issue on "Robust Statistical Techniques in Image Understanding." He is coauthor of an award winning paper in *Pattern Recognition* in 1989, the best student paper in 1999, and the best paper in the 2000 IEEE Conference Computer Vision and Pattern Recognition. His research interest is in application of modern statistical methods to image understanding problems. He is a senior member of the IEEE.

► For more information on this or any other computing topic, please visit our Digital Library at <http://computer.org/publications/dlib>.