Mean Shift: A Robust Approach toward Feature Space Analysis

Dorin Comaniciu[1] Peter Meer[2]

[1]Siemens Corporate Research
755 College Road East, Princeton, NJ 08540
comanici@scr.siemens.com

[2]Electrical and Computer Engineering Department
Rutgers University
94 Brett Road, Piscataway, NJ 08854-8058
meer@caip.rutgers.edu

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 equivalence 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.

Keywords: mean shift; clustering; image segmentation; image smoothing; feature space; low-level vision

1 Introduction

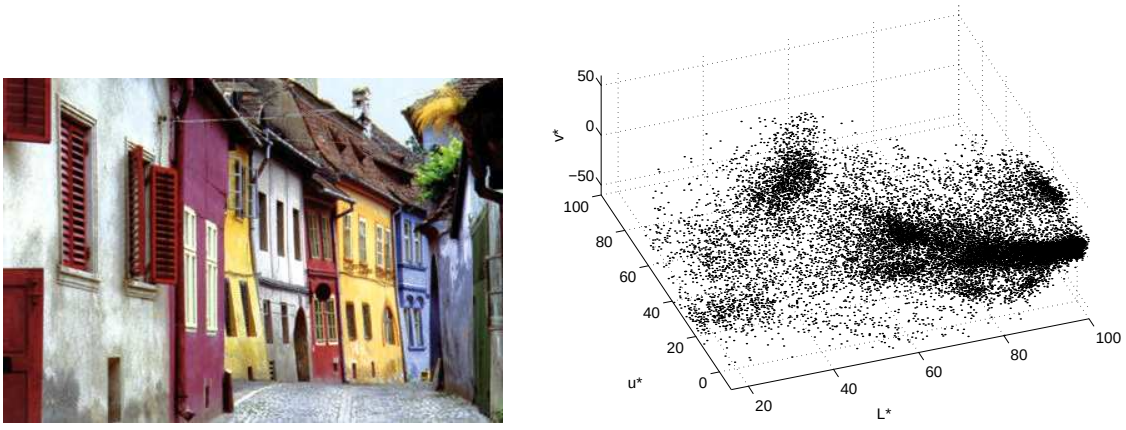
Low-level computer vision tasks are misleadingly difficult. Incorrect results can be easily obtained since the employed techniques often rely upon the user correctly guessing the values for the tuning parameters. To improve performance the execution of low-level tasks should be task driven, i.e., supported 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 is 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 are arising 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 an 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, Sec.8].



(a) (b)

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

In Figure 1 a typical example is shown. The color image in Figure 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 some proximity measure. See [28, Sec.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 nonparametric 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 till Cheng's paper [7] rekindled the interest in it. In spite of its excellent qualities, the mean shift procedure does not seem to be known in the statistical literature. While the book [54, Sec.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 implementational 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 the pattern recognition literature [17, Sec.4.3]) is the most popular density estimation method. Given n data points $\mathbf{x}_i$, $i = 1, \ldots, n$ 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

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

where

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

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

$$\begin{aligned} \int_{R^d} K(\mathbf{x}) \mathrm{d}\mathbf{x} = 1 \qquad & \lim_{\|\mathbf{x}\| \to \infty} \|\mathbf{x}\|^d K(\mathbf{x}) = 0 \\ \int_{R^d} \mathbf{x} K(\mathbf{x}) \mathrm{d}\mathbf{x} = 0 \qquad & \int_{R^d} \mathbf{x}\mathbf{x}^\top K(\mathbf{x}) \mathrm{d}\mathbf{x} = c_K \mathbf{I} \end{aligned} \tag{3}$$

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

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

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 $a_{k,d}^{-1} = \int_{R^d} K_1(\|\mathbf{x}\|) \mathrm{d}\mathbf{x}$ 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

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

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 $K(\mathbf{x})$ to 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 $\mathbf{H} = \mathbf{diag}[h_1^2, \ldots, h_d^2]$, 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

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

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 $n \to \infty$ while the bandwidth $h \to 0$ 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

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

which yields the radially symmetric kernel

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

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

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

yields the multivariate normal kernel

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

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

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

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)

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

We define the function

$$g(x) = -k'(x) \tag{13}$$

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

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

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

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

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

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

The second term is the *mean shift*

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

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) the expression (15) becomes

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

yielding

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

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

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

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, $\left\{\hat{f}_{h,K}(j)\right\}_{j=1,2...}$, is given by

$$\hat{f}_{h,K}(j) = \hat{f}_{h,K}(\mathbf{y}_j) \qquad j = 1, 2 \ldots . \tag{21}$$

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

Theorem 1 *If the kernel K has a convex and monotonically decreasing profile, the sequences $\{\mathbf{y}_j\}_{j=1,2...}$ and $\left\{\hat{f}_{h,K}(j)\right\}_{j=1,2...}$ converge, and $\left\{\hat{f}_{h,K}(j)\right\}_{j=1,2...}$ is also 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, Sec.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, Sec.1.2], moving in the direction of the local gradient guarantees convergence only for infinitesimal steps. The stepsize 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 stepsize 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

chose the adequate stepsizes. 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 $\left\{\mathbf{y}_j\right\}_{j=1,2...}$ and $\left\{\hat{f}_{h,K}(j)\right\}_{j=1,2...}$, 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

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

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

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

due to (19). Hence, $\mathbf{y}_c$ is a stationary point of $\hat{f}_{h,K}$.

Second, since $\left\{\hat{f}_{h,K}(j)\right\}_{j=1,2...}$ 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

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

The following theorem holds true for all $j = 1, 2, \ldots$, 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.,*

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

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 the relation (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, Chap.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 $f(x)$, $x_i = X_i,\ i = 1, \ldots, n$, while the relation between Z_i and X_i is

$$Z_i = m(X_i) + \epsilon_i \qquad i = 1, \ldots, n \tag{26}$$

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

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, Sec.3.1], [62, Sec.5.2], that the estimated constant is the value of the *Nadaraya–Watson* estimator

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

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

$$\mathrm{E}[\ (\hat{m}(x; h) - m(x)) \mid X_1, \ldots, X_n\ ] \approx h^2 \frac{m^{''}(x) f(x) + 2m'(x) f^{'}(x)}{2f(x)} \mu_2[g] \tag{28}$$

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

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

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

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

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

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

where $w(u) = \dfrac{d\rho(u)}{du}$. Therefore the iterations to find the location M-estimate are based on

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

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

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

which can be also interpreted as

$$\hat{\boldsymbol{\theta}} = \operatorname*{argmax}_{\theta} \hat{f}_{h,K}(\boldsymbol{\theta} \mid \mathbf{x}_1, \ldots, \mathbf{x}_n). \tag{34}$$

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.

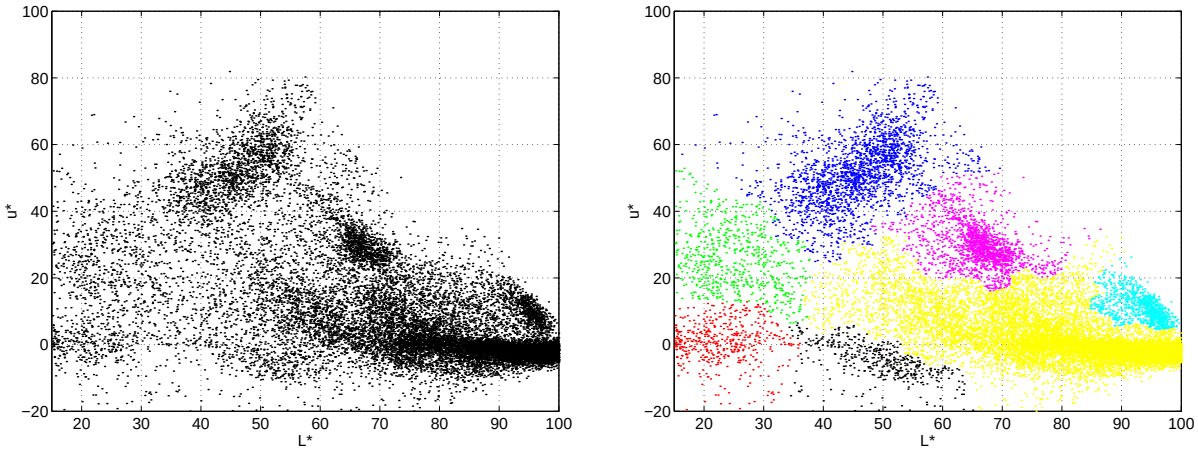
Before the analysis is performed two important (and somewhat related) issues should be addressed: the metric of the feature space and the shape of the kernel. The mapping from the input domain into a feature space often associates a noneuclidean metric to the space. The problem of color representation 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 terminate the iterations. Similarly, on flat plateaus the value of the gradient is close to zero and the mean shift procedure could stop.

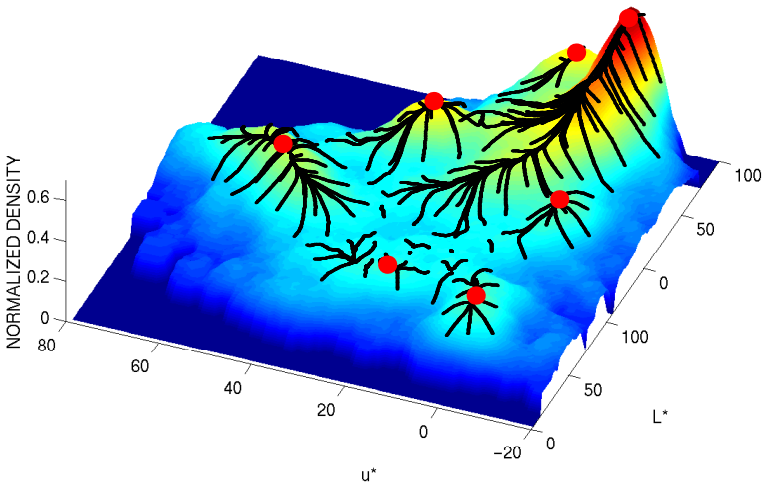
These artifacts are easy to eliminate through postprocessing. Mode candidates at a distance less than the kernel bandwidth are fused, the one corresponding to the highest density being chosen. The global structure of the feature space can be confirmed by measuring the significance of the valleys defined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



(a)

(b)



(c)

Figure 2: Example of a 2D feature space analysis. (a) Two dimensional data set of $110,400$ points representing the first two components of the $\mathrm{L^*u^*v^*}$ space shown in Figure 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.

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 Figure 2. A two dimensional data set of $110,400$ points (Figure 2a) is decomposed into 7 clusters represented with different colors in Figure 2b. A number of 159 mean shift procedures with uniform kernel were employed. Their trajectories are shown in Figure 2c, overlapped over the density estimate computed with 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 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 requires first the resampling of the input data with a regular grid. This 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 to design 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, Secs.6.5.2; 8.4]. The spaces $\mathrm{L^*u^*v^*}$ and $\mathrm{L^*a^*b^*}$ 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, Sec.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 $\mathrm{L^*u^*v^*}$ or $\mathrm{L^*a^*b^*}$, in the proposed algorithms we employed $\mathrm{L^*u^*v^*}$ 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 $\mathrm{L^*u^*v^*}$ 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, 3 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 an 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

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

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 nor-

malization 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.

Mean Shift Filtering

Let $\mathbf{x}_i$ and $\mathbf{z}_i,\ i = 1, \ldots, n$, 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 upperscripts 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, Sec.8E]. A similar (at least in spirit) twofold weighting is employed in the bilateral and mean shift based filterings, which is the main reason of 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 (Figure 3a), the result being shown in Figure 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 (600Mhz 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.

(a) (b)

Figure 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 Figure 3a is represented in three dimensions in Figure 4a. Note that the data was reflected over the horizontal axis of the window for a more informative display. In Figure 4b the mean shift paths associated with every other pixel (in both directions) 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 (Figure 4c) shows clean quasi-homogeneous regions.

The physical interpretation of the mean shift based filtering is easy to see by examining Figure 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.

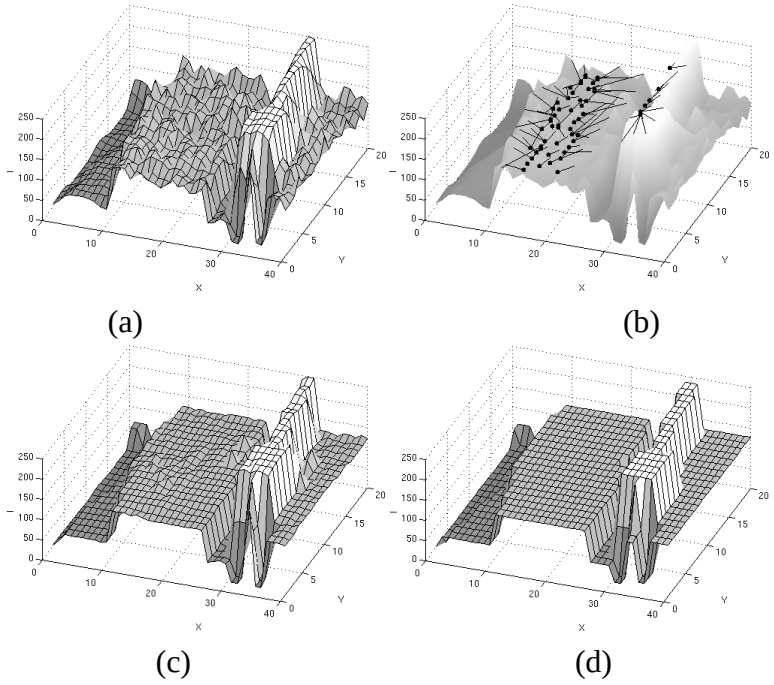


Figure 4: Visualization of mean shift based filtering and segmentation for gray level data. (a) Input. (b) Mean shift paths for the pixels on the plateau and on the line. The black dots are the points of convergence. (c) Filtering result $(h_s, h_r) = (8, 4)$. (d) Segmentation result.

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

Mean Shift Segmentation

Let $\mathbf{x}_i$ and $\mathbf{z}_i,\ i = 1, \ldots, n$, 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 $\{\mathbf{C}_p\}_{p=1\ldots m}$ 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

Original $(h_s, h_r) = (8, 8)$ $(h_s, h_r) = (8, 16)$

$(h_s, h_r) = (16, 4)$ $(h_s, h_r) = (16, 8)$ $(h_s, h_r) = (16, 16)$

$(h_s, h_r) = (32, 4)$ $(h_s, h_r) = (32, 8)$ $(h_s, h_r) = (32, 16)$

Figure 5: *Baboon* image. Original and filtered.

of attraction of the corresponding convergence points.

3. For each $i = 1, \ldots, n$, assign $L_i = \{p \mid \mathbf{z}_i \in \mathbf{C}_p\}$.
4. Optional: Eliminate spatial regions containing less than M pixels.

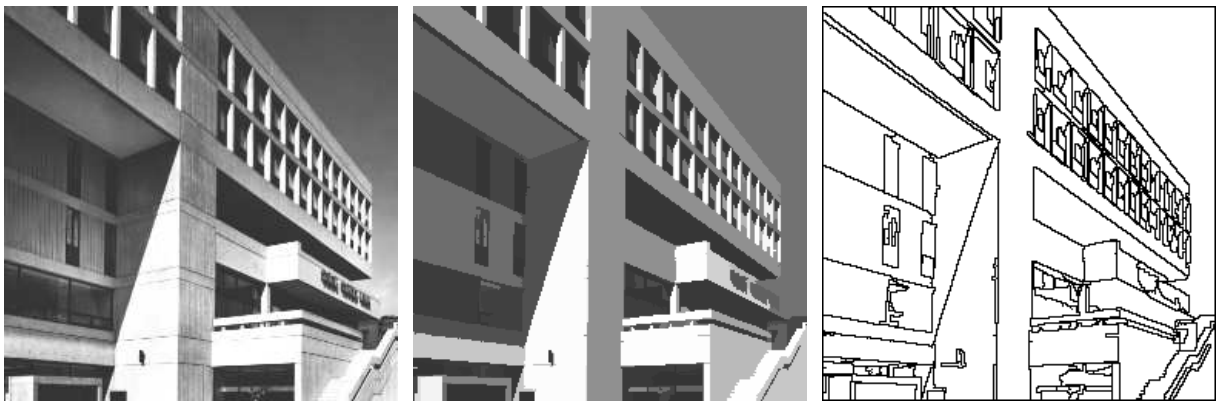
The cluster delineation step can be refined according to a priori information, and thus physics-based segmentation algorithms, e.g., [2, 35] can be incorporated. Since this process is performed on region adjacency graphs, hierarchical techniques like [36] can provide significant speed-up. The effect of the cluster delineation step is shown in Figure 4d. Note the fusion into larger homogeneous regions of the result of filtering shown in Figure 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 2000 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, Figure 6] with a similarly challenging example processed with the mean shift method [12, Figure 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 Figure 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, Figure 4] or by using a Gibbs random fields based approach [40, Figure 7].

The joint domain segmentation of the color 256×256 *room* image presented in Figure 7 is also satisfactory. Compare this result with the segmentation presented in [38, Figures 3e and 5c]



(a) (b) (c)

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

obtained by recursive thresholding. In both these examples, one can notice that regions in which a small gradient of illumination exist (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 to overcome the inherent limitations of methods based only on gray-level or color clustering, which typically oversegment small gradient regions.

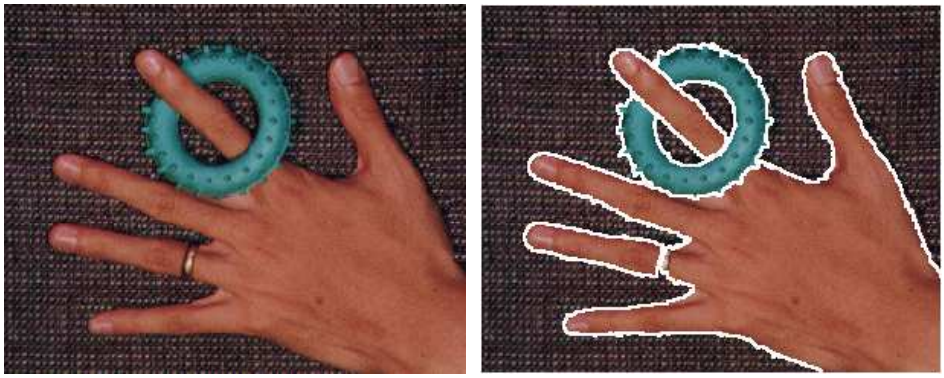
(a) (b)

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

The segmentation with $(h_s, h_r, M) = (16, 7, 40)$ of the 512×512 color image *lake* is shown in Figure 8. Compare this result with that of the multiscale approach in [57, Figure 11]. Finally, one can compare the contours of the color image $(h_s, h_r, M) = (16, 19, 40)$ *hand* presented in Figure 9 with those from [66, Figure 15] obtained through a complex global optimization, and

(a) (b)

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



(a) (b)

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

from [41, Figure 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

Figure 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.

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 Figure 10 and in which the original and labeled images compared in Figure 11.

As potential application of the segmentation, we return to the *cameraman* image. Figure 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 Figure 12b.

The (unoptimized) JAVA 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`

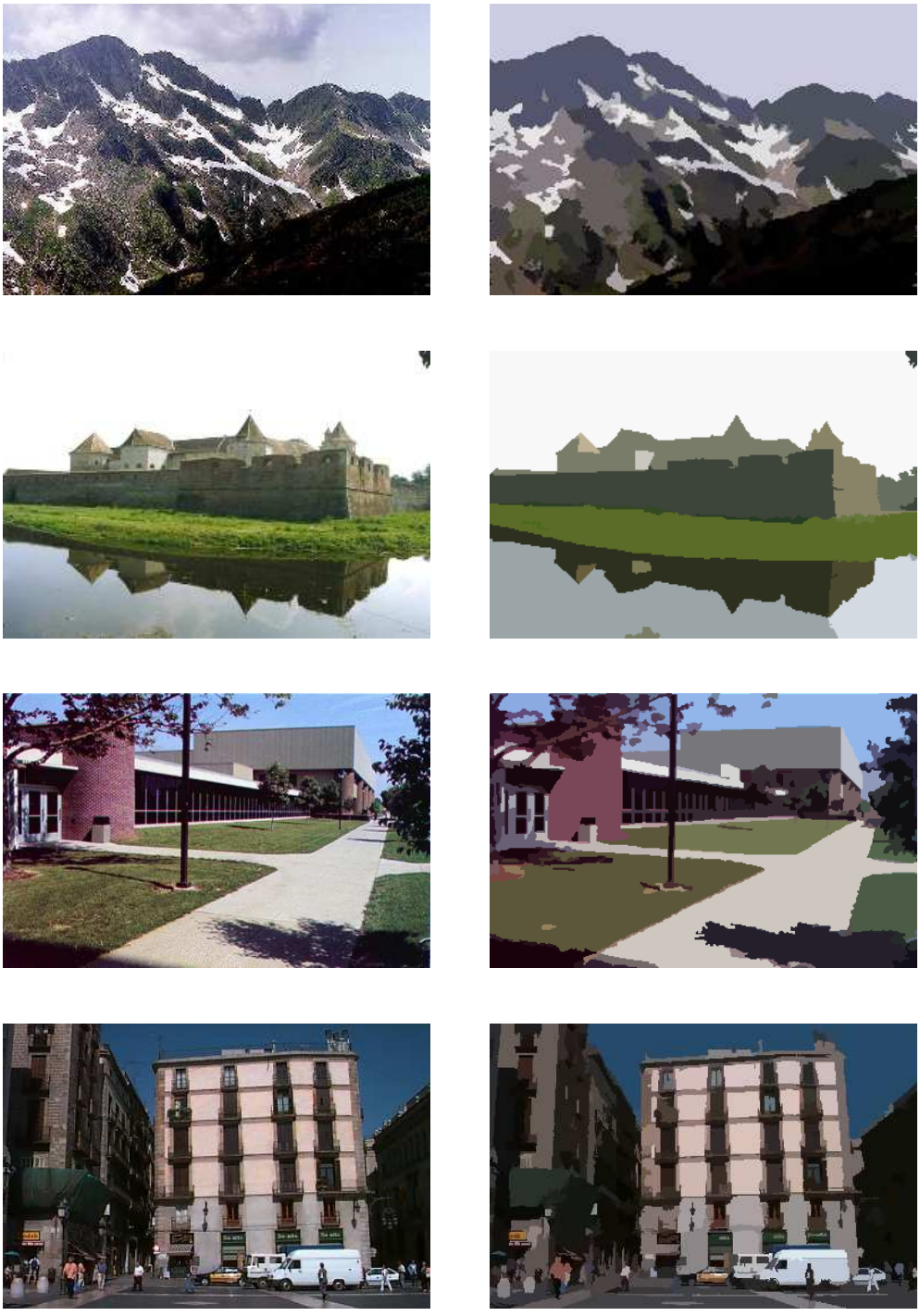


Figure 11: Some other segmentation examples with $(h_s, h_r, M) = (8, 7, 20)$. Left: original. Right: segmented.

(a) (b)

Figure 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.

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] already discusses a simple example. However, by introducing adequate objective functions the optimiza-

tion 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, non-rigid 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.

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 $\{\mathbf{y}_j\}_{j=1,2...}$ and $\left\{\hat{f}_{h,K}(j)\right\}_{j=1,2...}$ converge, and $\left\{\hat{f}_{h,K}(j)\right\}_{j=1,2...}$ is also 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 $j = 1, 2 \ldots$. Without loss of generality can be assumed that $\mathbf{y}_j = \mathbf{0}$ and thus from (16) and (21)

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

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

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

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

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

Using now (A.1) and (A.3) we obtain

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

and recalling (20) yields

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

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 $\mathbf{y}_{j+1} \neq \mathbf{y}_j = \mathbf{0}$, 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 $\left\{ \hat{f}_{h,K}(j) \right\}_{j=1,2\ldots}$ is convergent.

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

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

Summing now the two terms of the inequality (A.6) for indices $j, j+1 \ldots j+m-1$ it results that

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

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 $\left\{\mathbf{y}_j\right\}_{j=1,2\ldots}$.

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

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 w.l.g. 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

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

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 $\left(\mathbf{y}_{j+1} - \mathbf{x}_i\right)$ onto $\mathbf{y}_{j+1}$ is strictly negative, i.e.,

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

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

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

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

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

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

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

Introducing now (B.4) in (B.6) we have

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

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

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

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 have 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

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

or by taking into account (24)

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

Acknowledgment

The support of the National Science Foundation under the 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]. We would like to thank John Kent from University of Leeds and Dave 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," in *Proceedings of 2000 IEEE Conference on Computer Vision and Pattern Recognition*, Hilton Head, SC, volume II, June 2000, pp. 255–261.

[2] R. Bajcsy, S. W. Lee, and A. Leonardis, "Detection of diffuse and specular interface reflections and inter-reflections by color image segmentation," *International J. of Computer Vision*, vol. 17, pp. 241–272, 1996.

[3] D. Barash, "Bilateral filtering and anisotropic diffusion: Towards a unified viewpoint," Technical Report HPL-2000-18(R.1), Hewlett-Packard, 2000. Available at `www.hpl.hp.com/techreports`.

[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 Process.*, vol. 7, pp. 421–432, 1998.

[6] G. R. Bradski, "Computer vision face tracking as a component of a perceptual user interface," in *Proceedings IEEE Workshop on Applications of Computer Vision*, Princeton, NJ, October 1998, pp. 214–219.

[7] Y. Cheng, "Mean shift, mode seeking, and clustering," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 17, pp. 790–799, 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. Godtliebsen, and J. S. Maron, "Edge-preserving smoothers for image processing," *J. of Amer. Stat. Assoc.*, vol. 93, pp. 526–541, 1998.

[10] D. Comaniciu, *Nonparametric Robust Methods for Computer Vision*. PhD thesis, Department of Electrical and Computer Engineering, Rutgers University, 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," in *Proceedings of 1997 IEEE Conference on Computer Vision and Pattern Recognition*, San Juan, PR, June 1997, pp. 750–755.

[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," in *Proceedings 7th International Conference on Computer Vision*, Kerkyra, Greece, September 1999, pp. 1197–1203.

[14] D. Comaniciu, V. Ramesh, and P. Meer, "Real-time tracking of non-rigid objects using mean shift," in *Proceedings of 2000 IEEE Conference on Computer Vision and Pattern Recognition*, Hilton Head, SC, volume II, June 2000, pp. 142–149.

[15] D. Comaniciu, V. Ramesh, and P. Meer, "The variable bandwidth mean shift and data-driven scale selection," in *Proceedings 8th International Conference on Computer Vision*, Vancouver, Canada, volume I, July 2001, pp. 438–445.

[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," in *Proceedings 6th European Conference on Computer Vision*, Dublin, Ireland, volume II, June 2000, pp. 751–767.

[19] B. Fischl and E. L. Schwartz, "Adaptive nonlocal filtering: A fast alternative to anisotropic diffusion for image enhancement," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 21, pp. 42–48, 1999.

[20] K. Fukunaga, *Introduction to Statistical Pattern Recognition*. Academic Press, second edition, 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," in *SPIE Visual Communication and Image Processing*, San Jose, CA, volume 3653, 1999, pp. 1210–1221.

[23] G. Guy and G. Medioni, "Inference of surfaces, 3d curves, and junctions from sparse, noisy, 3d data," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 19, pp. 1265–1277, 1997.

[24] W. Härdle, *Applied Nonparameteric Regression*. Cambridge University 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 Recog. Letters*, vol. 17, pp. 1141–1150, 1996.

[26] P. J. Huber, *Robust Statistical Procedures*. SIAM, second edition, 1996.

[27] J. N. Hwang, 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 Anal. Machine Intell.*, vol. 22, pp. 4–37, 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 Anal. Machine Intell.*, vol. 7, pp. 165–177, 1985.

[32] G. Li, "Robust regression," in D. C. Hoaglin, F. Mosteller, and J. W. Tukey, editors, *Exploring Data Tables, Trends, and Shapes*, Wiley, 1985, pp. 281–343.

[33] L. Liu and S. Sclaroff, "Deformable shape detection and description via model-based region grouping," in *Proceedings of 1999 IEEE Conference on Computer Vision and Pattern Recognition*, Fort Collins, Colorado, volume II, June 1999, pp. 21–27.

[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 Anal. Machine Intell.*, vol. 13, pp. 307–316, 1991.

[37] J. Monteil and A. Beghdadi, "A new interpretation and improvement of nonlinear anisotropic diffusion for image enhancement," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 21, pp. 940–946, 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 Process.*, vol. 40, pp. 901–914, 1992.

[41] N. Paragios and R. Deriche, "Geodesic active contours for supervised texture segmentation," in *Proceedings of 1999 IEEE Conference on Computer Vision and Pattern Recognition*, Fort Collins, Colorado, volume II, June 1999, pp. 422–427.

[42] B. Park and J. Marron, "Comparison of data-driven bandwidth selectors," *J. of Amer. Stat. 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 Anal. Machine Intell.*, vol. 12, pp. 629–639, 1990.

[45] K. Popat and R. W. Picard, "Cluster-based probability model and its application to image and texture processing," *IEEE Trans. Image Process.*, vol. 6, pp. 268–284, 1997.

[46] W. K. Pratt, *Digital Image Processing*. Wiley, second edition, 1991.

[47] D. Ridder, J. Kittler, O. Lemmers, and R. Duin, "The adaptive subspace map for texture segmentation," in *Proceedings of 2000 International Conference on Pattern Recognition*, Barcelona, Spain, September 2000, pp. 216–220.

[48] T. Risse, "Hough transform for line recognition: Complexity of evidence accumulation and cluster detection," *Comput. Vision Graphics Image Process.*, vol. 46, pp. 327–345, 1989.

[49] S. J. Roberts, "Parametric and non-parametric unsupervised cluster analysis," *Pattern Recog.*, vol. 30, pp. 261–272, 1997.

[50] P. Saint-Marc, J. S. Chen, and G. Medioni, "Adaptive smoothing: A general tool for early vision," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 13, pp. 514–529, 1991.

[51] D. W. Scott, *Multivariate Density Estimation*. Wiley, 1992.

[52] R. Sedgewick, *Algorithms in C++*. Addison-Wesley, 1992.

[53] S. Sheather and M. Jones, "A reliable data-based bandwidth selection method for kernel density estimation," *J. Royal Statist. Soc. B*, vol. 53, pp. 683–690, 1991.

[54] B. W. Silverman, *Density Estimation for Statistics and Data Analysis*. Chapman & 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 Process.*, vol. 7, March 1998.

[57] M. Tabb and N. Ahuja, "Multiscale image segmentation by integrated edge and region detection," *IEEE Trans. Image Process.*, vol. 6, pp. 642–655, 1997.

[58] C. K. Tang, G. Medioni, and M. S. Lee, "Epipolar geometry estimation by tensor voting in 8d," in *Proceedings 7th International Conference on Computer Vision*, Kerkyra, Greece, volume I, September 1999, pp. 502–509.

[59] C. Tomasi and R. Manduchi, "Bilateral filtering for gray and color images," in *Proceedings 6th International Conference on Computer Vision*, Bombay, India, January 1998, pp. 839–846.

[60] A. Touzani and J. G. Postaire, "Clustering by mode boundary detection," *Pattern Recog. 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," in *Proceedings 6th International Conference on Computer Vision*, Bombay, India, January 1998, pp. 67–72.

[62] M. P. Wand and M. Jones, *Kernel Smoothing*. Chapman & Hall, 1995.

[63] R. Wilson and M. Spann, "A new approach to clustering," *Pattern Recog.*, 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 Anal. Machine Intell.*, vol. 19, pp. 780–785, 1997.

[65] G. Wyszecki and W. S. Stiles, *Color Science: Concepts and Methods, Quantitative Data and Formulae*. Wiley, second edition, 1982.

[66] S. C. Zhu and A. Yuille, "Region competition: Unifying snakes, region growing, and bayes/mdl for multiband image segmentation," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 18, pp. 884–900, 1996.

[67] X. Zhuang, Y. Huang, K. Palaniappan, and Y. Zhao, "Gaussian mixture density modeling decomposition, and applications," *IEEE Trans. Image Process.*, vol. 5, pp. 1293–1302, 1996.