A Robust Probabilistic Estimation Framework for Parametric Image Models*

Maneesh Singh, Himanshu Arora, and Narendra Ahuja

University of Illinois at Urbana-Champaign, Urbana 61801, USA,
{msingh,harora1,n-ahuja}@uiuc.edu,
http://vision.ai.uiuc.edu/~msingh

Abstract. Models of spatial variation in images are central to a large number of low-level computer vision problems including segmentation, registration, and 3D structure detection. Often, images are represented using parametric models to characterize (noise-free) image variation, and, additive noise. However, the noise model may be unknown and parametric models may only be valid on individual segments of the image. Consequently, we model noise using a nonparametric kernel density estimation framework and use a locally or globally linear parametric model to represent the noise-free image pattern. This results in a novel, robust, redescending, M- parameter estimator for the above image model which we call the Kernel Maximum Likelihood estimator (KML). We also provide a provably convergent, iterative algorithm for the resultant optimization problem. The estimation framework is empirically validated on synthetic data and applied to the task of range image segmentation.

1 Introduction

Models of spatial variation in visual images are central to a large number of low level computer vision problems. For example, segmentation requires statistical models of variation within individual segments. Registration requires such models for finding correspondence across images. Image smoothing depends on distinction between bona fide image variation versus noise. Achieving super-resolution critically depends on the nature of photometric variation underlying an image of a given resolution. A general image model consists of two components - one that captures noise-free image variation, and the other that models additive noise. Robust solutions to low level vision problems require specification of these components. Local or piecewise variation can often be modelled in a simple parametric framework. However, recognizing that a model is applicable only in a specific neighborhood of a data point is a challenge. Added to this is the complex and data dependent nature of noise. Robust estimation of model parameters must address both of these challenges. As concrete examples of these challenges, (a) consider the problem of estimating the number of planar patches

* The support of the Office of Naval Research under grant N00014-03-1-0107 is gratefully acknowledged.

T. Pajdla and J. Matas (Eds.): ECCV 2004, LNCS 3021, pp. 508–522, 2004.

in a range image. One cannot compute a least squares planar fit for the entire range image data as there might be several planes[1]. (b) Further, range data may be corrupted with unknown additive noise. In this paper we use linear parametric models to represent local image variation and develop a robust framework for parameter estimation.

Much work has been done on robust parameter estimation in statistics and more recently in vision and it is difficult to refer to all of it here. We point the reader to some important papers in this area. The seminal paper by Huber [1] defined M-estimators (maximum-likelihood like estimators) and studied their asymptotic properties for linear models. This analysis is carried forward by Yohai and Maronna [2], and Koenker and Portnoy [3], among others. Recently, Chu et al. [4] introduced a redescending M-estimator for robust smoothing that preserves image edges (though the estimator is not consistent). Hillebrand and Müller [5] modified this estimator and proved its consistency. In computer vision, M-estimation has been extensively used for video coding [6], optical flow estimation [7], pose estimation [8], extraction of geometric primitives from images [9] and segmentation [10]. For a review on robust parameter estimation in computer vision, refer to Stewart [11]. More recently, the problem of robust estimation of local image structure have been extensively studied by [12,13,14], however they implicitly assume a locally constant image model while our model is more general.

Our main contributions are: (A) We show that a redescending M-estimator for linear parametric models can be derived through Hyndman et al.'s [15] modification to kernel regression estimators that reduces its bias to zero. This provides a link between nonparametric density estimation and parametric regression. It is illustrative to compare our approach with Chen and Meer [16], who use a geometric analogy to derive the estimator whereas we use the maximum likelihood approach. (B) The solution to the M-estimator (under mild conditions) is a nonlinear program (NLP). We provide a fast, iterative algorithm for the NLP and prove its convergence properties. (C) We show the utility of our algorithm for the task of segmentation of piecewise-planar range images.

In Section 2, we define the problem of robust linear regression. In Section 3, we first present the likelihood framework for parameter estimation using zero bias-in-mean kernel density estimators and then we provide a solution to the estimation problem. In Section 4, we empirically evaluate the performance of the proposed estimator and compare it with the least squares and least trimmed squares estimators under various noise contaminations. Finally, in Section 5, we show an instance of a vision problem where the algorithm can be applied.

[1] One may point out the use of the Hough transform for this task. However, the Hough transform is a limiting case of the robust parameter estimation model that we develop here.

2 Linear Regression

In [17], we presented a kernel density estimation framework for the image model,

$$Y(\mathbf{t}) = I(\mathbf{t}) + \epsilon(\mathbf{t}) \tag{1}$$

where $Y(\cdot)$, the observed image, is a sum of the clean image signal $I : \mathcal{R}^d \to \mathcal{R}$, and a noise process $\epsilon : \mathcal{R}^d \to \mathcal{R}$. In this paper, we use a linear parametric model for the image signal. More specifically, $I(\cdot)$ is either locally or globally linear in model parameters. We assume that the noise is unknown. This relaxation in assumptions about noise allows us to develop robust models for parameter estimation.

Definition 1 (Linear Parametric Signal Model) *The signal model in (1) is called linear in the model parameters,* $\Theta \doteq [\theta_1, \theta_2, \cdots, \theta_d]^T$*, if the function* $I(\mathbf{t})$ *can be expressed as* $I(\mathbf{t}) \doteq \mathbf{g}(\mathbf{t})^T\Theta = \sum_{i=1}^{d} \theta_i g_i(\mathbf{t})$ *for some* $\mathbf{g}(\mathbf{t}) \doteq [g_1(\mathbf{t}), \ldots, g_d(\mathbf{t})]$*.*

The problem of linear regression estimation is to estimate the model parameters $\Theta = [\theta_1, \ldots, \theta_d]^T$ given a set of observations[2] $\{\mathbf{z}_i \doteq [Y_i; \mathbf{t}_i]\}_{i=1}^n$. One popular framework is to use least squares (LS) estimation where the estimate minimizes the mean square error between the sample values and those predicted by the parametric model.

Definition 2 (Least Squares Estimate) *Given a sample* $\{\mathbf{z}_i\}_{i=1}^n$ *and a fixed, convex set of weights* $\{w_i\}_{i=1}^n$*, the least weighted squares estimate* $\hat{\Theta}_{LWS}$ *for the linear parametric image model in Definition 1 is given by,* $\hat{\Theta}_{LWS} = \min_{\Theta \in \mathcal{R}^d} \sum_{i=1}^n w_i \epsilon_i^2(\Theta)$ *where* $\epsilon_i(\Theta) \doteq Y_i - \mathbf{g}(\mathbf{t}_i)^T\Theta$*. When the weights are all equal, then* $\hat{\Theta}_{LWS}$ *is the standard least squares estimate, denoted by* $\hat{\Theta}_{LS}$*.*

It is well known that the LS estimate is the maximal likelihood estimate if the errors, $\{\epsilon_i\}$, are zero-mean, independent and identically distributed with a normal distribution. In practice, however, the error distribution is not normal and often unknown. Since the error pdf is required for optimal parameter estimation and is often unknown, robust estimators have been used to discard some observations as outliers [11]. In this paper we formulate a maximum likelihood estimation approach based on the estimation of error pdf using kernel density estimators. Such an approach yields a robust redescending M-estimator. Our approach, however, has two advantages: (1) The formulation has a simple, convergent, iterative solution, and, (2) information from other sources (for example, priors on the parameters) can be factored in via a probabilistic framework.

[2] We use $[a; b]$ to denote a vertical concatenation of column vectors, i.e., $[a^T, b^T]^T$

3 Kernel Estimators for Parametric Regression

A conditional density estimator for image signals using probability kernels can be defined as follows (see [17] for properties and details),

Definition 3 (Kernel Conditional Density Estimator) *Let* $\mathbf{z} = [Y; \mathbf{t}] = [Y, x_1, \ldots, x_d]^T \in \mathcal{R}^d$ *be a (d+1)-tuple. Then, we write the kernel estimator for the conditional density of* Y *given the spatial location* $\mathbf{t} = [x_1, \ldots, x_d]^T$ *as,*

$$\hat{f}_{Y|\mathbf{t}}(v|\mathbf{u}) = \frac{1}{n|H|\mathcal{D}} \sum_{i=1}^{n} K(H^{-1}[v - Y(\mathbf{t}_i); (\mathbf{u} - \mathbf{t_i})]) \tag{2}$$

where H *is a non-singular* $(d+1) \times (d+1)$ *bandwidth matrix and* $K : \mathcal{R}^{d+1} \to \mathcal{R}$ *is a kernel such that it is non-negative, has a unit area (*$\int_{\mathcal{R}^{d+1}} K(\mathbf{z})\, d\mathbf{z} = 1$*), zero mean (*$\int_{\mathcal{R}^{d+1}} \mathbf{z}K(\mathbf{z})\, d\mathbf{z} = 0$*), and, unit covariance (*$\int_{\mathcal{R}^{d+1}} \mathbf{z}\mathbf{z}^T K(\mathbf{z})\, d\mathbf{z} = I_{d+1}$*).*

For data defined on a regular grid, such as for images, $\mathcal{D} = \frac{1}{n|H|} \sum_{i=1}^{n} \int_{\mathcal{R}} K(H^{-1}(\mathbf{z}_i - \mathbf{z}))\, dY =: \frac{1}{n} \sum_{i=1}^{n} w_i(\mathbf{t})$ can be treated as a normalization constant which clearly does not depend upon the values $\{\mathbf{Y}_i\}_i$. $w_i(\mathbf{t})$ or simply w_i (when the spatial location $\mathbf{t}$ is unambiguous) is given by $\frac{1}{|H|} \int_{\mathcal{R}} K(H^{-1}(\mathbf{z}_i - \mathbf{z}))\, dY$. We sometimes assume that the probability kernels are separable or rotationally symmetric or both. In the sequel, we require a specific kind of separability: we would like the spatial domain and the intensity domain to be separable. We shall also assume that the kernels are rotationally symmetric with a convex profile (please refer to Chapter 6, [18]).

3.1 Zero Bias-in-Mean Kernel Regression

Thus, we will assume kernel $K(\cdot)$ to be separable in the following sense: Let $\mathbf{z} \doteq [Y, \mathbf{t}^T]^T]$ denote any data point such that the first coordinate is its intensity component and the rest are its spatial coordinates. Kernel K is separable in the subspaces represented by the intensity and spatial coordinates, denoted by $\mathcal{D}_Y$ and $\mathcal{D}_{\mathbf{t}} \doteq \mathcal{R}^d \ominus \mathcal{D}_Y$. Assuming $H^{-1} = \begin{pmatrix} 1/h_Y & 0_{1\times d} \\ 0_{d\times 1} & H_{\mathbf{t}}^{-1} \end{pmatrix}$, $K(H^{-1}\mathbf{z}) = K_Y(\frac{Y}{h_Y}) K_{\mathbf{t}}(H_{\mathbf{t}}^{-1}\mathbf{t})$.

In (2), values $\{Y_i \doteq Y(\mathbf{t}_i)\}_i$ at locations $\{\mathbf{t}_i\}_i$ are used to *predict* the behavior of Y at location $\mathbf{u}$. For the kernel estimate, the expected value of Y conditioned on $\mathbf{u}$ is given by, $E[Y|\mathbf{u}] = \sum_{i=1}^{n} w_i(\mathbf{u})$. Since conditional mean is independent of the regression model, we need to factor the model in Definition 1, into the conditional kernel density estimate in (2). The necessity for doing so is explained using Figure 1(a). At location $\mathbf{u} = 5$ shown on the horizontal axis, $E[Y|\mathbf{u}]$ will have a positive bias since most values, Y_i, in the neighborhood are larger than the *true* intensity value at $\mathbf{u} = 5$. The conditional mean will have this bias if the image is not locally linear. In general, the conditional density estimate will be biased wherever the image is not (locally) constant. This effect was noted

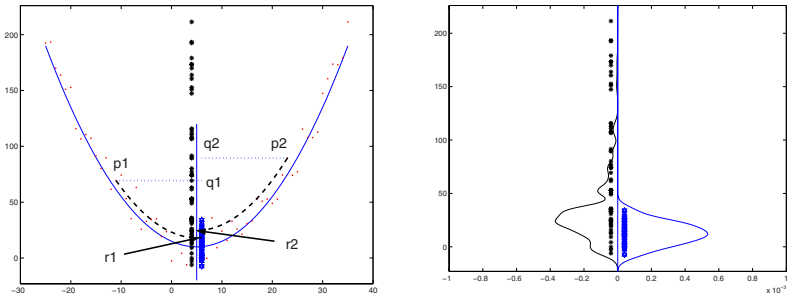


Fig. 1. This figure shows that kernel density estimation after factoring in the image model leads to better estimates. (a) Shown are the data points generated using the 1D image model $Y(t) = 10 + 0.2 * (t-5)^2 + N(0, 10^2)$. To estimate the density $f(Y|t=5)$ according to Definitions 3 and 4, data points $p1$ and $p2$ are moved along straight and curved lines respectively. (b) Shown are the two estimated pdfs, $\hat{f}_{Y|\mathbf{t}}$ and $\tilde{f}_{Y|\mathbf{t}}$ to the left and right respectively ($h_t = 10$,$h_Y = 5$).

by Hyndman et. al. [15]. To address this defect, they proposed the following modified kernel estimator which we call the zero bias-in-mean kernel (0-BIMK) conditional density estimator.

Definition 4 (0-BIMK Conditional Density Estimator) *The zero bias-in-mean kernel (0-BIMK) estimator for the conditional pdf is given by,* $\tilde{f}_{Y|\mathbf{t}}(v|\mathbf{u}) \doteq \frac{1}{n|H|\mathcal{D}} \sum_{i=1}^{n} K(H^{-1}[v - \tilde{Y}(\mathbf{t}_i); (\mathbf{u} - \mathbf{t_i})])$ *such that,* $\tilde{Y}_i = I(\mathbf{u}) + \epsilon_i - \overline{\epsilon}$, $\epsilon_i = Y_i - I(\mathbf{t}_i)$, $\overline{\epsilon} = \sum_{i=1}^{n} w_i \epsilon_i$.

The conditional mean of Y using the above density estimator is, $\tilde{E}[Y|\mathbf{x}] = \sum_{i=1}^{n} w_i \tilde{Y}_i = I(\mathbf{x})$. Thus, the conditional mean of Y computed using the 0-BIMK estimator is guaranteed to be that predicted by the model since the conditional density at any domain point is computed by factoring in the assumed underlying model. In Figure 1(a), we show how the density estimate is constructed using points p1 and p2. While intensity values q1 and q2 (after spatial weighing) are used to find $\hat{f}_{Y|\mathbf{t}}$ at $t = 5$, r1 and r2 are used to estimate $\tilde{f}_{Y|\mathbf{t}}$. The two estimates are depicted in Figure 1(b) - among them, the latter estimate is clearly a *better* estimate of $10 + N(0, 10^2)$.

3.2 0-BIM Parametric Kernel Regression

We are interested in parametric regression models where the regression model, $I(t)$, has the form $I(\mathbf{t}) = \Theta^T \mathbf{g}(\mathbf{t})$. Our goal is to estimate the model parameters, Θ and thus, the regression function, $I(\mathbf{t})$. Let the parameter estimate be $\hat{\Theta}$. Then, the regression estimate is given by $\hat{I}(\mathbf{t}) = \hat{\Theta}^T \mathbf{g}(\mathbf{t})$. Using Definition 4, the kernel density estimator using the parametric regression model is defined below.

Definition 5 (0-BIMK Parametric Density Estimator) *The zero bias-in-mean kernel (0-BIMK) conditional density estimator, using the parametric regression model $I(\mathbf{t}) = \Theta^T \mathbf{g}(\mathbf{t})$ in Definition 1, is given by,*

$$\tilde{f}_{Y|\mathbf{t}}(v|\mathbf{u}, \Theta) \doteq \frac{1}{n|H|\mathcal{D}} \sum_{i=1}^{n} K(H^{-1}[v - \tilde{Y}(\mathbf{t}_i); (\mathbf{u} - \mathbf{t_i})]) \tag{3}$$

with $\tilde{Y}_i \doteq g(\mathbf{u})^T\Theta + \epsilon_i(\Theta) - \overline{\epsilon}(\Theta)$, $\epsilon_i \doteq Y_i - g(\mathbf{t}_i)^T\Theta$ and $\overline{\epsilon}(\Theta) \doteq \sum_{j=1}^{n} w_j \epsilon_j(\Theta)$.

The minimum variance parameter estimator (KMV) for the above density estimate is not a particularly good choice. It is easy to see that the KMV estimate chooses that pdf for the noise process which minimizes the estimated noise variance. This is clearly undesirable if the error pdf is multi-modal. In such a scenario, kernel pdf estimation becomes useful. It provides us with the knowledge that the data is multimodal and assists in choosing the correct mode for the parameter estimate.[3]. This prompts the following variant of the ML parameter estimation procedure (one might define other such criteria).

Definition 6 *Kernel Maximum Likelihood Parameter Estimator, $\tilde{\Theta}_{KML}$ or simply, $\tilde{\Theta}$, is defined to be that value of Θ which maximizes the likelihood that $Y(\mathbf{t}) = I(\mathbf{t})$ when the likelihood is estimated using the 0-BIMK estimator.*

$$\tilde{\Theta}_{KML} \doteq \max_{\Theta \in \mathcal{R}^d} \tilde{f}_{Y|\mathbf{t}}(I(\mathbf{t})|\mathbf{t}, \Theta) \tag{4}$$

3.3 Estimation of $\tilde{\Theta}$

$\tilde{\Theta}$ is the solution of an unconstrained and differentiable nonlinear program (NLP) which requires global optimization techniques. We introduce now an iterative nonlinear maximization algorithm that is guaranteed to converge to a local maximum for any starting point. The main attractiveness of our algorithm is that it chooses the step size automatically. The convergence result holds if the probability kernel, $K(.)$, has a convex profile (please refer to Chapter 6, [18]). The algorithm can be repeated for several starting points for global maximization (a standard technique for global optimization). We build the proof of convergence using the following two results.

Theorem 7 (Least Squares Fit) *Given a data sample $\{[Y_i, \mathbf{t}_i^T]\}_{i=1}^n$, the Least Squares fit of the model $Y \approx \Theta^T g(\mathbf{t})$, defined by $\hat{\Theta}_{LWS}$ in Definition 2 is a solution of the normal equations, $\sum_{i=1}^{n} w_i g(\mathbf{t}_i) g(\mathbf{t}_i)^T \tilde{\Theta}_{LS} = \sum_{i=1}^{n} w_i g(\mathbf{t}_i) Y_i$.*

[3] The cost of this generalization achieved by kernel pdf estimation, however, is that one needs to know the scale of the underlying pdf at which it is to be estimated from its samples.

Lemma 8 *Let there be n points,* $\{\mathbf{t}_i\}_{i=1}^n$*, in* $\mathcal{R}^d$*, m functions,* $g_i : \mathcal{R}^d \rightarrow \mathcal{R}$*,* $i = 1, \cdots, m$ *and strictly positive convex weights* $\{w_i\}_{i=1}^n$*. Then the matrix* $A \doteq \sum_{i=1}^n w_i(g(\mathbf{t}_i) - \overline{g(\mathbf{t})})(g(\mathbf{t}_i) - \overline{g(\mathbf{t})})^T$ *is invertible if and only if there exist* n' *distinct points,* $n \geq n' \geq m$*, such that the set of functions,* $\{\sqrt{w_i}(g(\mathbf{t}_i) - \overline{g(\mathbf{t})})\}_{i=1}^m$*, is independent over them.*

Proof. Let Q_A be the quadratic form associated with A. For every $\mathbf{y} \in \mathcal{R}^d$, $Q_A(\mathbf{y}) \doteq \mathbf{y}^T A \mathbf{y} \geq 0$ since $Q_A(\mathbf{y}) = \sum_{i=1}^n w_i((g(\mathbf{t}_i) - \overline{g(\mathbf{t})}) \cdot y)^2 \geq 0$. Thus, A is positive semidefinite. Hence A is invertible iff it is positive definite. Further, A is positive definite iff the set of vectors $\{\sqrt{w_i}(g(\mathbf{t}_i) - \overline{g(\mathbf{t})})\}_i$ spans $\mathcal{R}^d$. Evidently, this is possible iff there exist n' distinct points, such that $n \geq n' \geq m$ and the set of functions $\{\sqrt{w_i}(g(\mathbf{t}_i) - \overline{g(\mathbf{t})})\}_{k=1}^m$ are independent over such a set. □

Theorem 9 (Parametric Mean Shift using Weighted Least Squares) *Let there be n points,* $\{\mathbf{t}_i\}_{i=1}^n$*, in* $\mathcal{R}^d$ *such that at least m points are distinct. Further, let there be m independent functions,* $g_i : \mathcal{R}^d \rightarrow \mathcal{R}$*,* $i = 1, \cdots, m$ *as in Definition 6. Also, let K be such that* $K(H^{-1}\mathbf{z}) = K_Y(\frac{Y}{h_Y})K_{\mathbf{t}}(H_{\mathbf{t}}^{-1}\mathbf{t})$*. If* K_Y *has a convex and non-decreasing profile* κ*, then the sequence* $\{\tilde{f}(j)\}_{j=1,2,\ldots} \doteq \{\tilde{f}_{Y|\mathbf{t}}(I(\mathbf{t})|\mathbf{t}, \Theta_j)\}_{j=1,2,\ldots}$*, of probability values computed at* $\mathbf{z} = (I(\mathbf{t}), \mathbf{t})$ *with corresponding parameter estimates* $\{\Theta_j\}_{j=1,2,\ldots}$ *defined*[4] *by* $\Theta_{j+1} = \text{argzero}_\Theta A_j(\Theta - \Theta_j) - b_j$ *according to (7) below, is convergent. The sequence* $\{\Theta_j\}_{j=1,2,\ldots}$ *converges to a local maximum of* $\{\tilde{f}_{Y|\mathbf{t}}(I(\mathbf{t})|\mathbf{t}, \Theta)\}$*.*

Proof. From Definition (5),

$$\tilde{f}(\Theta_j) \doteq \tilde{f}_{Y|\mathbf{t}}(I(\mathbf{t})|\mathbf{t}, \Theta_j) = \frac{1}{n|H|\mathcal{D}} \sum_{i=1}^n K(H^{-1}[\epsilon_i(\Theta_j) - \bar{\epsilon}(\Theta_j); (\mathbf{u} - \mathbf{t_i})])$$

$$= \frac{1}{h_Y} \sum_{i=1}^n K_Y(\frac{\epsilon_i(\Theta_j) - \bar{\epsilon}(\Theta_j)}{h_Y}) w_i$$

Using convexity of the profile κ, defining $w_{i,j} = \kappa'(-\frac{(\epsilon_i(\Theta_j) - \bar{\epsilon}(\Theta_j))^2}{h_Y^2}) w_i$ and denoting all the multiplicative constants by C, we get[5],

$$\tilde{f}(\Theta_{j+1}) - \tilde{f}(\Theta_j) \geq C \sum_{i=1}^n w_{i,j}\{(\epsilon_{i,j}^c)^2 - (\epsilon_{i,j}^c - (\Theta_{j+1} - \Theta_j)^T(g(\mathbf{t}_i) - \overline{g(\mathbf{t})}))^2\}$$

$$\text{(where, } \epsilon_{i,j}^c \doteq (Y_i - \bar{Y}) - \Theta_j^T(g(\mathbf{t}_i) - \overline{g(\mathbf{t})}) \text{ and } \overline{g(t)} \doteq \sum_i w_i g(t)) \tag{5}$$

The above equation shows that the RHS and consequently, the LHS is non-negative if solution to the following LWS problem exists.

$$\Theta_{j+1} - \Theta_j \doteq \underset{\Theta \in \mathcal{R}^d}{\text{argmin}} \sum_{i=1}^n w_{i,j}(\epsilon_{i,j}^c - \Theta^T(g(\mathbf{t}_i) - \overline{g(\mathbf{t})}))^2$$

[4] $\text{y} = \text{argzero}_x\, g(x)$ denotes a value of y such that $g(y) = 0$

[5] Since $\tilde{f}(\Theta_{j+1}) - \tilde{f}(\Theta_j) \geq C \sum_{i=1}^n w_{i,j}((\epsilon_i(\Theta_j) - \bar{\epsilon}(\Theta_j))^2 - (\epsilon_i(\Theta_{j+1}) - \bar{\epsilon}(\Theta_{j+1}))^2) = C \sum_{i=1}^n w_{i,j}(((Y_i - \bar{Y}) - \Theta_j^T(g(\mathbf{t}_i) - \overline{g(\mathbf{t})}))^2 - ((Y_i - \bar{Y}) - \Theta_{j+1}^T(g(\mathbf{t}_i) - \overline{g(\mathbf{t})}))^2)$

From Theorem (7), $\Theta_{j+1} - \Theta_j$ must satisfy the following normal equations,

$$\sum_{i=1}^{n} w_{i,j}(g(\mathbf{t}_i) - \overline{g(\mathbf{t})})(g(\mathbf{t}_i) - \overline{g(\mathbf{t})})^T(\Theta_{j+1} - \Theta_j)$$
$$= \sum_{i=1}^{n} w_{i,j}\epsilon^c_{i,j}(g(\mathbf{t}_i) - \overline{g(\mathbf{t})}) = \nabla_\Theta \tilde{f}(\Theta_j) \quad (6)$$

which has the form $A_j(\Theta_{j+1} - \Theta_j) = b_j$ where

$$A_j \doteq \sum_{i=1}^{n} w_{i,j}(g(\mathbf{t}_i) - \overline{g(\mathbf{t})})(g(\mathbf{t}_i) - \overline{g(\mathbf{t})})^T; b_j \doteq \sum_{i=1}^{n} w_{i,j}\epsilon^c_{i,j}(g(\mathbf{t}_i) - \overline{g(\mathbf{t})}) \quad (7)$$

Further, Lemma (8) guarantees the existence of a solution. Hence, RHS of (5)

$$= C\sum_{i=1}^{n} w_{i,j}\{2\epsilon^c_{i,j}(\Theta_{j+1} - \Theta_j)^T(g(\mathbf{t}_i) - \overline{g(\mathbf{t})}) - \{(\Theta_{j+1} - \Theta_j)^T(g(\mathbf{t}_i) - \overline{g(\mathbf{t})})\}^2\}$$
$$= C\sum_{i=1}^{n} w_{i,j}\{(\Theta_{j+1} - \Theta_j)^T(g(\mathbf{t}_i) - \overline{g(\mathbf{t})})\}^2 = (\nabla_\Theta \tilde{f}(j))^T(\Theta_{j+1} - \Theta_j) \geq 0$$

where the equalities follow from (6). The above equation shows that $S_f \doteq \{\tilde{f}(\Theta_j)\}_j$ is a nondecreasing sequence. Since the sequence is obviously bounded, it is convergent. RHS of the third equality above also shows that the solution to the normal equations always produces a step which has a positive projection along the gradient of the estimated pdf. By Lemma (8), $A_j, j = 1, 2, \ldots$ are fully ranked for all $j's$. RHS of the second equality above also implies that $\|\delta\Theta_j\| \to 0$ where $\delta\Theta_j \doteq \Theta_{j+1} - \Theta_j$.

Now, we show that the sequence $S_\Theta \doteq \{\Theta_j\}_{j=1,2,\ldots}$ is convergent[6]. As the sequence is bounded, it has an isolated (since n is finite) limit point Θ^*. Since $\delta\Theta_j \to 0$, given small enough r, ϵ, $0 < r < r + \epsilon$ and open balls $B(\Theta^*, r)$, $B(\Theta^*, r + \epsilon)$, there exists an index J_1 such that for all $j > J_1$, $\|\delta\Theta_j\| < \epsilon$, Θ^* is the only limit point in $B(\Theta^*, r + \epsilon)$ and the set $U \doteq \{\Theta | \Theta \in B(\Theta^*, r + \epsilon) \cap B^c(\Theta^*, r), \Theta \in S\}$ in non-empty. U has finite items, let their maximum value be M_U. Further, as S_f is strictly increasing, there exists an index J_2 such that for all $j > J_2$, $\tilde{f}(\Theta_j) > M_U$. We define $J = \max(J_1, J_2)$. Then, for any $j > J$, $\Theta_j \in B(\Theta^*, r) \Rightarrow \Theta_{j+1} \in B(\Theta^*, r)$ since $B(\Theta^*, r) \cup B(\Theta_j, \delta\Theta_j) \subset B(\Theta^*, r + \epsilon)$ and $\Theta_{j+1} \notin U$. Further, $\delta\Theta_j \to 0 \Rightarrow b_j = \nabla_\Theta \tilde{f}(\Theta_j) \to 0$. Hence, Θ^* is a point of local maximum for $\tilde{f}(\cdot)$. □

Thus, Theorem (4) guarantees a solution to the kernel maximum likelihood problem in the sense that given any starting point, we would converge to a local maximum of the parametric kernel likelihood function. However, at each iteration step, a weighted least squares problem needs to be solved that requires matrix

[6] We gratefully acknowledge Dr. Dorin Comaniciu's assistance with the proof.

inversion. Further, it is required that in the limit, the matrix A_j stays positive definite. This condition need not always be satisfied. In the next theorem, we provide a way to bypass this condition.

Theorem 10 (Parametric Mean Shift using Gradient Descent) *Let the-re be n points, $\{\mathbf{t}_i\}_{i=1}^n$, in $\mathcal{R}^d$ such that at least m points are distinct. Further, let there be m independent functions, $g_i : \mathcal{R}^d \to \mathcal{R}$, $i = 1, \cdots, m$ in terms of which we want to find $\tilde{\Theta}_{KML}$ as given in Definition 6. Also, let K be such that $K(H^{-1}\mathbf{z}) = K_Y(\frac{Y}{h_Y})K_{\mathbf{t}}(H_{\mathbf{t}}^{-1}\mathbf{t})$. If K_Y has a convex and non-decreasing profile κ, then the sequence $\{\tilde{f}(j)\}_{j=1,2,\ldots} \doteq \{\tilde{f}_{Y|\mathbf{t}}(I(\mathbf{t})|\mathbf{t}, \Theta_j)\}_{j=1,2,\ldots}$ of probability values computed at $\mathbf{z} = (I(\mathbf{t}), \mathbf{t})$ with corresponding parameter estimates $\{\Theta\}_{j=1,2,\ldots}$, defined by $\Theta_{j+1} = \Theta_j + k_j \nabla_\Theta \tilde{f}(\Theta_j)$ in (8), is convergent. The sequence $\{\Theta_j\}_{j=1,2,\ldots}$ converges to a local maximum of $\{\tilde{f}_{Y|\mathbf{t}}(I(\mathbf{t})|\mathbf{t}, \Theta)\}$.*

Proof. From (5),

$$\tilde{f}(\Theta_{j+1}) - \tilde{f}(\Theta_j) \geq C \sum_{i=1}^{n} w_{i,j}(\epsilon_{i,j}^c)^2 - (\epsilon_{i,j}^c - (\Theta_{j+1} - \Theta_j)^T(g(\mathbf{t}_i) - \overline{g(\mathbf{t})}))^2$$

Now, we take the next step in the direction of the gradient, i.e., we seek some $k_j > 0$ such that $\Theta_{j+1} = \Theta_j + k_j \nabla_\Theta \tilde{f}(\Theta_j)$ and k_j is the solution of,

$$k_j = \underset{k}{\operatorname{argmin}} \sum_{i=1}^{n} w_{i,j}(\epsilon_{i,j}^c - k \sum_{l=1}^{n} w_{l,j}\epsilon_{l,j}^c g^c(t_l)^T g^c(t_i))^2$$

Solving for k_j, we get,

$$k_j = \frac{\| \sum_{i=1}^n w_{i,j}\epsilon_{i,j}^c g(\mathbf{t}_i)\|^2}{\sum_{i=1}^n w_{i,j}\langle g(\mathbf{t}_i), \sum_{l=1}^n w_{l,j}\epsilon_{l,j}^c g(\mathbf{t}_l)\rangle^2} =: \frac{N(k_j)}{D(k_j)} \geq \frac{1}{\sum_{i=1}^n w_{i,j}\|g(\mathbf{t}_i)\|^2} > 0$$

Hence,

$$\text{RHS of (8)} = C \sum_{i=1}^{n} w_{i,j} \left[{\epsilon_{i,j}^c}^2 - (\epsilon_{i,j}^c - k_j \sum_{l=1}^{n} w_{l,j}\epsilon_{i,j}^c g(\mathbf{t}_l)^T g(\mathbf{t}_l))^2 \right]$$
$$= C\frac{N(k_j)\|\nabla_\Theta \tilde{f}(\Theta_j)\|^2}{D(k_j)} = C(\Theta_{j+1} - \Theta_j) \cdot \nabla_\Theta f(\Theta_j) = \frac{C}{k_j}\|\Theta_{j+1} - \Theta_j\|^2 \geq 0$$

This implies the convergence of $S_f \doteq \{\tilde{f}(\Theta_j)\}_j$ and that $\|\delta\Theta_j\| \to 0$ as in the proof for Theorem 4. The convergence of $S_\Theta \doteq \{\Theta_j\}_{j=1,2,\ldots}$ to a local maximum similarly follows. □

The gradient descent algorithm is more stable (albeit slower) since it does not require matrix inversion. Hence, we use this algorithm for applications in Section 5.

4 Empirical Validation

We tested the robustness of our algorithm empirically and compared its statistical performance with (a) the least squares (LS) method, and, (b) the least trimmed squares (LTS) method. The comparison to LS was done since it is the most popular parameter estimation framework apart from being the minimum-variance unbiased estimator (MVUE) for additive white Gaussian noise (AWGN), the most commonly used noise model. Further, our algorithm is a re-weighted LS (RLS) algorithm. LTS [19] was chosen since it is a state-of-the-art method for robust regression. We note here that LS is not robust while LTS has a positive breakdown value of ([(n-p)/2] + 1)/n [19]. In computer vision, one requires robust algorithms that have a higher breakdown value [20]. Redescending M-estimators (which the proposed formulation yields) are shown to have such a property. Thus, theoretically, they are more robust then either of the estimators described above. Nonetheless, it is useful to benchmark the performance of the proposed estimator with the above two estimators.

For ease of comparison, we used a 1-D data set as our test bed (Table 1). Intensity samples were generated using the equation $I_0(x) = ax^2 + bx + c$ at unit intervals from $x = -50$ to $x = 50$. To test the noise performance, we added uncorrelated noise from a noise distribution to each intensity sample independently. We used Gaussian and two-sided log-normal distributions as the noise density functions. We tested the performance at several values of the variance parameter associated with these distributions. At each value of variance, we generated 1000 realizations and calculated the bias and the variance of the two estimators. The Gaussian distribution was chosen as the LS estimator is the minimum variance unbiased estimator (MVUE) for AWGN noise while the log-normal distribution was chosen to simulate outliers. We present the results in Table 1: results for i.i.d. Gaussian noise are presented in the left column and for i.i.d. log-normal noise are presented in the right column. The ground-truth values are $a = 0.135$, $b = 0.55$ and $c = 1.9$. The results show that the performance of the proposed estimator is better than the LTS and the LS estimators when the number of outliers are large (log-normal noise). In case the noise is AWGN, standard deviation of the proposed estimator stays within twice that of the LS estimator (which is the MVUE). However, the proposed algorithm has a better breakdown value than either of the two estimators . Further, the proposed estimator is easy to implement as a recursive least squares (RLS) algorithm, and, in our simulations, it was an order of magnitude faster than the LTS algorithm.

5 Applications

The theoretical formulation derived in the previous sections provides a robust mechanism for parametric (global) or semi-parametric (local) image regression. This is achieved via a parametric representation of the image signal coupled with non-parametric kernel density estimation for the additive noise. As a consequence, the proposed estimation framework can be used for the following tasks:

Table 1. LS, LTS and KML estimates for (a, b, c). $I_i = ax_i^2 + bx_i + c + \epsilon_i$. $\{\epsilon_i\}$ are i.i.d. Gaussian and i.i.d. log-normal with parameters $(0, \sigma^2)$ for experiments in left and right columns respectively. Ground-truth values are $(a, b, c) = (0.135, 0.55, 1.9)$. Mean and standard deviation of the estimated values for 1000 experiments are presented for each estimator - LS, LTS and KML, in rows 1, 2, and 3, respectively. KML performs better for log-normal while LTS and LS perform better for gaussian noise.

LS - Gaussian noise

	Mean			Standard Deviation		
σ	a	b	c	a	b	c
1	0.135	0.550	1.907	0.0001	0.0036	0.1516
5	0.135	0.549	1.952	0.0007	0.0165	0.7474
9	0.135	0.549	1.853	0.0012	0.0308	1.3658
13	0.135	0.548	1.951	0.0017	0.0449	1.9375

(a)

LS - log-normal noise

	Mean			Standard Deviation		
σ	a	b	c	a	b	c
1	0.135	0.550	1.917	0.0004	0.0098	0.4016
2	0.135	0.555	1.729	0.0063	0.1369	7.6999
3	0.140	0.194	6.673	0.2862	6.7165	2.36e2
4	0.762	1.605	-8.7e2	2.00e1	3.68e2	3.05e4

(b)

LTS - Gaussian noise

	Mean			Standard Deviation		
λ	a	b	c	a	b	c
1	0.135	0.549	1.905	0.0002	0.0036	0.1881
5	0.135	0.547	2.005	0.0007	0.0177	0.8838
9	0.135	0.551	2.000	0.0011	0.0292	1.2843
13	0.135	0.550	2.003	0.0015	0.0335	1.7416

(c)

LTS - log-normal noise

	Mean			Standard Deviation		
λ	a	b	c	a	b	c
1	0.135	0.549	1.810	0.0005	0.0148	0.6224
2	0.135	0.550	1.996	0.0007	0.0179	0.8286
3	0.135	0.549	1.963	0.0009	0.0257	1.0629
4	0.135	0.548	1.973	0.0014	0.0380	1.7173

(d)

KML - Gaussian noise

	Mean			Standard Deviation		
σ	a	b	c	a	b	c
1	0.135	0.549	1.904	0.0002	0.0036	0.2697
5	0.135	0.547	1.960	0.0008	0.0166	1.0517
9	0.135	0.546	1.950	0.0015	0.0311	1.8904
13	0.135	0.544	1.951	0.0021	0.0481	2.6815

(e)

KML - log-normal noise

	Mean			Standard Deviation		
σ	a	b	c	a	b	c
1	0.135	0.549	1.931	0.0004	0.0083	0.4570
2	0.135	0.548	1.927	0.0007	0.0164	0.7446
3	0.135	0.548	1.916	0.0007	0.0198	0.8184
4	0.135	0.547	1.992	0.0011	0.0204	1.5543

(f)

(A1) Edge preserving denoising of images that can be modelled (locally or globally) using the linear parametric model in Definition 1. (A2) Further, since the KML Estimation framework admits a multimodal error model, the framework can be used to partition data generated from a mixture of sources, each source generating data using the aforementioned parametric model (in such a case, we use $H_{\mathbf{t}} \to \infty$. In other words, spatial kernel is assumed identically equal to one). Thus, the proposed framework provides a systematic way of doing Hough Transforms. Indeed, Hough Transform is the discretized version of $\tilde{f}_{Y|\mathbf{t}}(\Theta^T g(\mathbf{t}))$ if $H_Y \to 0$ and $H_{\mathbf{t}} \to 0$.

To illustrate an application of the proposed estimation framework, we apply it to the task of range image segmentation. As test images, we use the database[7] [21] generated using the ABW structured light camera. The objects imaged in

[7] Range Image Databases provided on the University of South Florida website http://marathon.csee.usf.edu/range/DataBase.html.

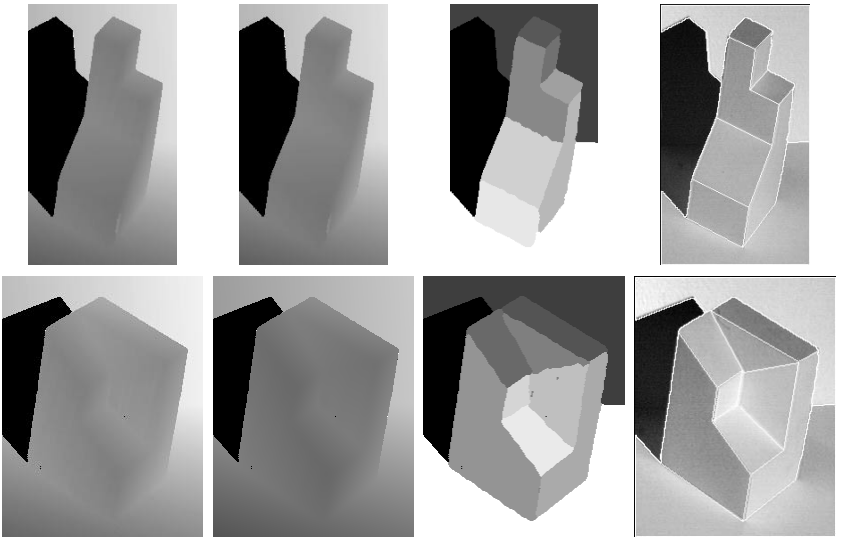


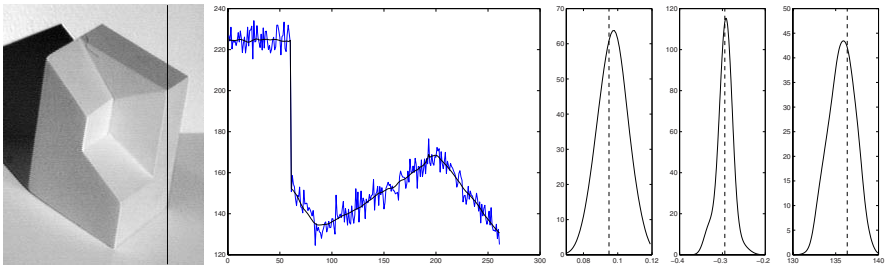
Fig. 2. Results for ABW test images 27 (row 1) and 29 (row 2). (a) Original range image, (b) reconstructed range image $\hat{I}(\mathbf{t})$ (locally planar assumption), (c) extracted planar segments, (d) Final segmentation after geometric boundary refinement. Edges are superimposed on the intensity images.

the database are all piecewise linear and provide a good testbed for the proposed estimation framework.

Let the range image be denoted by $Y(\mathbf{t})$. Then, for each $\mathbf{t}$ in the image domain $\mathcal{D}$, the range may be represented as $Y(\mathbf{t}) = \langle [\mathbf{t}^T 1]^T, \Theta(\mathbf{t}) \rangle + \epsilon(\mathbf{t})$ where $\Theta(\mathbf{t}) \in S_\Theta \doteq \{\Theta_1, \cdots, \Theta_M\}$ such that there are M unknown planar regions in the range image. The task of planar range image segmentation is to estimate the set S_Θ and to find the mapping, $\mathcal{T} : \mathcal{D} \rightarrow S_\Theta$.

The algorithm used for segmentation has following steps: (1) Estimate[8] local parameters, $\tilde{\Theta}(\mathbf{t})$ for each $\mathbf{t}$. This also provides an edge-preserved de-noised range image ($\hat{I}(\mathbf{t})$). (2) Sort $\tilde{\Theta}(\mathbf{t})$'s in descending order of likelihood values, $\tilde{f}_{Y|\mathbf{t}}(\langle [\mathbf{t}^T 1]^T, \tilde{\Theta}(\mathbf{t}) \rangle)$. (3) Starting with the most likely $\tilde{\Theta}(\mathbf{t})$, estimate the parameters (say, $\tilde{\Theta}_p$) of the largest (supported) plane (spatial bandwidth matrix ∞ for global estimation). Since kernel maximum likelihood function can have several local minima, sorting has the effect of avoiding these local minima. Next, remove the data points supported by this plane ($\tilde{f}_{Y|\mathbf{t}}(\langle [\mathbf{t}^T 1]^T, \tilde{\Theta}_p(\mathbf{t}) \rangle)$ greater than a threshold) from the set of data points as well as from the sorted list. (4) Choose the next (remaining) point from the sorted list and fit the next largest plane to the remaining data points. (5) Iterate Step (4) N times ($N >> M$, we choose $N = 100$) or until all data points get exhausted. (6) Discard the planes smaller than k pixels (resulting in N_1 planes),

[8] KML estimation requires specification of the bandwidth parameters. These are hand-selected currently although they can be estimated akin to [17].



(a) (b) (c)

Fig. 3. (a) ABW test image 29. (b) Noisy and reconstructed range image (locally planar assumption) values are shown, along the scan line in (a), depicting robustness to discontinuities. (c) Estimated histogram of computed planar parameters for each data point of the leftmost object plane are shown. True image model : $I = 0.095y - 0.294x + 136.35$.

and construct the set $S_\Theta \doteq \{\tilde{\Theta}_1, \cdots, \tilde{\Theta}_{N_1}\}$. (7) Reclassify all data points based on the likelihood values, i.e. the function $\mathcal{T}(\cdot)$ is initially estimated to be $\mathcal{T}(\mathbf{t}) = \text{argmax}_{i \in 1 \ldots N_1} \tilde{f}_{Y|\mathbf{t}}(\langle [\mathbf{t}^T 1]^T, \tilde{\Theta}_p(\mathbf{t}) \rangle \; | (\mathbf{t}), \tilde{\Theta}_i)$. (8) Repeat step (6). (9) Refine boun-daries based on the geometric reasoning that all intersecting planes produce straight lines as edges (refer to Chapter 6, [18]).

In Figure 2, we show segmentation results on two[9] range images from the ABW data set. The range values in column (a) properly belong to piecewise planar range data. The range data is missing in the (black) shadowed regions. We treat these regions as planar surfaces that occlude all neighboring surfaces. The estimated locally linear surface (output of step 1, $\hat{I}(\mathbf{t})$) is shown in column (b). To better illustrate the results of KML estimation, we reproduce a scan line in Figure 3(a). Note, in Figure 3(b), that occluding edge (discontinuity in the range data) is well preserved while denoising within continuous regions takes place as expected. This illustrates that the kernel maximum likelihood estimation preserves large discontinuities while estimating image parameters in a robust fashion. At the same time, locally estimated planar parameters are close to true values (refer to Figure 3(c) and the caption).

After planar regions are extracted in step (7), we show the results in Figure 2(c). Here, the brightness coding denotes different label numbers. We see that very few spurious regions are detected - this validates the robustness of global parameter estimation framework since we fit global planes to the data and iteratively remove data points supported by the planes. The goodness of the estimated plane parameters now allows us to carry out the geometric boundary refinement step as we can now estimate the corner and edge locations as detailed above. The final result after this is depicted in Figure 2(d). As one can see, the estimated segmentation is near perfect. To see that our results are better than several well-known range image segmentation algorithms, please compare them visually with those presented in Hoover et al [21].

[9] More results will be included in the two extra pages allowed for the final manuscript.

6 Conclusions

In this paper, we presented a robust maximum likelihood parameter estimation framework by modelling the image with a linear parametric model and additive noise using kernel density estimators. The resulting estimator, the KML Estimator, is a redescending M-estimator. This novel approach provides a link between robust estimation of parametric image models and nonparametric density models for additive noise. We also provided a solution to the resultant nonlinear optimization problem and proved its convergence. Finally, we apply our framework to range image segmentation.

References

1. Huber, P.J.: Robust regression: Asymptotics, conjectures and monte carlo. The Annals of Statistics **1** (1973) 799–821
2. Yohai, V.J., Maronna, R.A.: Asymptotic behavior of M-estimators for the linear model. The Annals of Statistics **7** (1979) 258–268
3. Koenker, R., Portnoy, S.: M-estimation of multivariate regressions. Journal of Amer. Stat. Assoc. **85** (1990) 1060–1068
4. Chu, C.K., Glad, I.K., Godtliebsen, F., Marron, J.S.: Edge-preserving smoothers for image processing. Journal of the American Statistical Association **93** (1998) 526–541
5. Hillebrand, M., Muller, C.H.: On consistency of redescending M-kernel smoothers. unpublished manuscript - http://www.member.uni-oldenburg.de/ch.mueller/publik/neuarticle5.pdf **9** (2000) 1897–1913
6. Ayer, S., Sawhney, H.: Layered representation of motion video using robust maximum likelihood estimation of mixture models and MDL encoding. Computer Vision, Int'l Conf. (1995) 777–784
7. Black, M.J., Jepson, A.D.: Estimating optical flow in segmented images using variable-order parametric models with local deformations. PAMI, IEEE Trans. **18** (1996) 972–986
8. Kumar, R., Hanson, A.R.: Robust methods of estimating pose and a sensitivity analysis. Image Understanding, CVGIP (1994)
9. Liu, L., Schunck, B.G., Meyer, C.C.: On robust edge detection. Robust Computer Vision, Int'l Workshop (1990) 261–286
10. Mirza, M.J., Boyer, K.L.: Performance evaluation of a class of M-estimators for surface parameter estimation in noisy range data. IEEE Trans., Robotics and Automation **9** (1993) 75–85
11. Stewart, C.V.: Robust parameter estimation in computer vision. SIAM Reviews **41** (1999) 513–537
12. Tomasi, C., Manduchi, R.: Bilateral filtering for gray and color images. Computer Vision, 6th Intl. Conf. (1998) 839–846
13. Barash, D., Comaniciu, D.: A common framework for nonlinear diffusion, adaptive smoothing, bilateral filtering and mean shift. Image and Video Computing (2003)
14. Boorngaard, R., Weijer, J.: On the equivalence of local-mode finding, robust estimation and mean-shift analysis as used in early vision tasks. Pattern Recognition, 16th Intl. Conf. **3** (2002) 927–930

15. Hyndman, R.J., Bashtannyk, D.M., Grunwald, G.K.: Estimating and vizualizing conditional densities. Journal of Comput. and Graph. Statistics **5** (1996) 315–336
16. Chen, H., Meer, P.: Robust computer vision through kernel density estimation. Computer Vision, 7th Euro. Conf. **1** (2002) 236–250
17. Singh, M.K., Ahuja, N.: Regression-based bandwidth selection for segmentation using Parzen windows. Computer Vision, 9th Intl. Conf. (2003) 2–9
18. Singh, M.K.: Image Segmentation and Robust Estimation Using Parzen Windows. University of Illinois at Urbana-Champaign. Ph.D. Thesis (2003)
19. Rousseeuw, P.J., Van Driessen, K.: Computing LTS Regression for Large Data Sets. University of Antwerp. Technical Report (1999)
20. Mizera, I., Muller, C.H.: Breakdown points and variation exponents of robust M-estimators in linear models. Annals of Statistics **27** (1999) 1164–1177
21. Hoover, A., J.-B., G., Jiang, X., Flynn, P.J., Bunke, H., Goldgof, D.B., Bowyer, K.K., Eggert, D.W., Fitzgibbon, A.W., Fisher, R.B.: An experimental comparison of range image segmentation algorithms. PAMI, IEEE Trans. **18** (1996) 673–689