# **Robust Regression Methods for Computer Vision: A Review**

PETER MEER,\* DORON MINTZ, AND AZRIEL ROSENFELD Center for Automation Research, University of Maryland, College Park, MD 20742

#### DONG YOON KIM

Agency for Defense Development, P.O. Box 35, Taejeon, Korea

## Abstract

Regression analysis (fitting a model to noisy data) is a basic technique in computer vision. Robust regression methods that remain reliable in the presence of various types of noise are therefore of considerable importance. We review several robust estimation techniques and describe in detail the least-median-of-squares (LMedS) method. The method yields the correct result even when half of the data is severely corrupted. Its efficiency in the presence of Gaussian noise can be improved by complementing it with a weighted least-squares-based procedure. The high time-complexity of the LMedS algorithm can be reduced by a Monte Carlo type speed-up technique. We discuss the relationship of LMedS with the RANSAC paradigm and its limitations in the presence of noise corrupting all the data, and we compare its performance with the class of robust M-estimators. References to published applications of robust techniques in computer vision are also given.

#### 1 Introduction

Regression analysis (fitting a model to noisy data) is an important statistical tool frequently employed in computer vision for a large variety of tasks. Tradition and ease of computation have made the least squares method the most popular form of regression analysis. The least squares method achieves optimum results when the underlying error distribution is Gaussian. However, the method becomes unreliable if the noise has nonzero-mean components and/or if outliers (samples with values far from the local trend) are present in the data. The outliers may be the result of clutter, large measurement errors, or impulse noise corrupting the data. At a transition between two homogeneous regions of the image, samples belonging to one region may become outliers for fits to the other region.

Three concepts are usually employed to evaluate a regression method: relative efficiency, breakdown point, and time complexity. The *relative efficiency* of a regression method is defined as the ratio between the lowest achievable variance for the estimated parameters (the Cramer-Rao bound) and the actual variance provided by the given method. The efficiency also depends

\*Present address: Department of Electrical and Computer Engineering, Rutgers University, Piscataway, NJ 08855-0909 U.S.A. on the underlying noise distribution. For example, in the presence of Gaussian noise the mean estimator has an asymptotic (large sample) efficiency of 1 (achieving the lower bound) while the median estimator's efficiency is only  $2/\pi = 0.637$  (Mosteller & Tukey 1977).

The *breakdown point* of a regression method is the smallest amount of outlier contamination that may force the value of the estimate outside an arbitrary range. For example, the (asymptotic) breakdown point of the mean is 0 since a single large outlier can corrupt the result. The median remains reliable if less than half of the data are contaminated, yielding asymptotically the maximum breakdown point, 0.5.

The *time complexity* of the least squares method is  $O(np^2)$  where *n* is the number of data points and *p* is the number of parameters to be estimated. Feasibility of the computation requires a time complexity of at most  $O(n^2)$ .

A new, improved regression method should provide:

- a. reliability in the presence of various types of noise, i.e., good asymptotic and small sample efficiency;
- b. protection against a high percentage of outliers, i.e., a high breakdown point;
- c. a time complexity not much greater than that of the least squares method.

Many statistical techniques have been proposed which satisfy some of the above conditions. These techniques are known as *robust regression* methods. In section 2 a review of robust regression methods is given. In section 3 the least-median-of-squares (LMedS) method is discussed in detail. A comparison of the LMedS method with the RANSAC paradigm in section 4 gives us the opportunity to analyze its behavior for data usually met in computer vision problems. Through a simple application of LMedS, windowoperator based smoothing, we compare its performance in section 5 with the class of robust M-estimators. The article is concluded in section 6.

#### 2. Robust Regression Methods

The early attempts to introduce robust regression methods involved straight-line fitting. In one class of methods the data is the first partitioned into two or three nearly equal sized parts ( $i \le L$ ;  $L < i \le R$ ; R < i), where *i* is the index of the data and L = R in the former case. The slope  $\beta_1$  and the intercept  $\beta_0$  of the line are found by solving the system of nonlinear equations

where med represents the median operator applied to the set defined below it. The breakdown point of the method is 0.5/k where k is the number of partitions (2 or 3) since the median is used for each part separately. Brown and Mood (1951) investigated the method for k = 2, and Tukey introduced the resistant line procedure for k = 3 (see Johnstone & Velleman 1985).

Another class of methods uses the slopes between each pair of data points without splitting up the data set. Theil (1950) estimated the slope as the median of all n(n - 1)/2 slopes which are defined by *n* data points. The breakdown point of these methods is 0.293 since at least half the slopes should be correct in order to obtain the correct estimate. That is, if  $\epsilon$  is the fraction of outliers in the data we must have  $(1 - \epsilon)^2 \ge 0.5$ . The intercept can be estimated from the input data by employing the traditional regression formula.

The theory of multidimensional robust estimators was developed in the seventies. The basic robust estimators are classified as M-estimators, R-estimators and Lestimators (Huber 1981).

The M-estimators are the most popular robust regression methods. (See the book by Hampel et al. (1986) for a leading reference.) These estimators minimize the sum of a symmetric, positive-definite function  $\rho(r_i)$  of the residuals  $r_i$ , with a unique minimum at  $r_i = 0$ . (A residual is defined as the difference between the data point and the fitted value.) For the least squares method  $\rho(r_i) = r_i^2$ . Several  $\rho$  functions have been proposed which reduce the influence of large residual values on the estimated fit. Huber (1981) employed the squared error for small residuals and the absolute error for large residuals. Andrews (1974) used a squared sine function for small and a constant for large residuals. Beaton and Tukey's (1974) biweight is another example of these  $\rho$ functions. The M-estimates of the parameters are obtained by converting the minimization

$$\min \sum_{i} \rho(r_i) \tag{2}$$

into a weighted least squares problem which then is handled by available subroutines. The weights depend on the assumed  $\rho$  function and the data. The reliability of the initial guess is of importance and the convergence of the solution is not proved for most of the  $\rho$  functions. Holland and Welsch (1977) developed algorithms for solving the numerical problems associated with M-estimators. We will return to the subject in section 5.

R-estimators are based on ordering the set of residuals. Jaeckel (1972) proposed obtaining the parameter estimates by solving the minimization problem

$$\min \sum_{i} a_n(R_i) r_i \tag{3}$$

where  $r_i$  is the residual;  $R_i$  is the location of the residual in the ordered list, that is, its rank; and  $a_n$  is a score function. The score function must be monotonic and

$$\sum_i a_n(R_i) = 0$$

The most frequently used score function is that of Wilcoxon:  $a_n(R_i) = R_i - (n + 1)/2$ . Since  $|a_n(R_i)| \le (n - 1)/2$ , the largest residuals caused by outliers cannot have too large a weight. Scale invariance (independence from the variance of the noise) is an important advantage of R-estimators over M-estimators. Cheng and Hettmansperger (1983) presented an iteratively

reweighted least squares algorithm for solving the minimization problem associated with R-estimators.

The L-estimators employ linear combinations of order statistics. The median and  $\alpha$ -trimmed-mean based methods belong to this class. It is important to notice, however, that the mean ( $\alpha = 0$ ) is a least squares estimate, while the median can be regarded also as the M-estimate obtained for  $\rho = |r_i|$ . Various simulation studies have shown that L-estimators give less satisfactory results than the other two classes (Heiler 1981).

In spite of their robustness for various distributions the M-, R- and L-estimators have breakdown points that are less than 1/(p + 1), where p is the number of parameters in the regression (Hampel et al. 1986, p. 329; Li 1985). For example, in planar surface fitting we have p = 3, and the breakdown point is less than 0.25, making it sensitive to outliers.

Recently several robust estimators having breakdown points close to 0.5 were proposed. Siegel (1982) introduced the repeated-median (RM) method of solving multidimensional regression problems. Suppose pparameters are to be estimated from n data samples. A parameter is estimated in the following way: First, for each possible p-tuple of samples the value of the parameter is computed yielding a list of C(n, p) (the binomial coefficient) terms. Then the medians for each of the p indexes characterizing a p-tuple are obtained recursively. When the list has collapsed into one term, the result is the RM estimate of the parameter. Once a parameter has been estimated, the amount of computation can be reduced for the remaining p - 1parameters.

For example, let p = 3 and suppose that we start by estimating the parameter  $\beta_2$  of the planar fit

$$z = \beta_0 + \beta_1 x + \beta_2 y \tag{4}$$

First the values

$$\beta_2(i,j,k) = \frac{(z_i - z_j)(x_j - x_k) - (z_j - z_k)(x_i - x_j)}{(y_i - y_j)(x_j - x_k) - (y_j - y_k)(x_i - x_j)}$$
(5)

are computed for all the triplets defined by  $i \neq j \neq k$ . The estimate is then

$$\hat{\beta}_2 = \underset{i}{\operatorname{med}} \underset{j(\neq i)}{\operatorname{med}} \underset{k(\neq i,j)}{\operatorname{med}} \beta_2(i, j, k) \tag{6}$$

The parameter  $\beta_1$  is estimated next, by applying the same algorithm for p = 2 to the data  $z_i - \hat{\beta}_2 y_i$ . Similarly the value of  $\beta_0$  is obtained by taking the median of the samples  $z_i - \hat{\beta}_2 y_i - \hat{\beta}_1 x_i$ . The breakdown point of the repeated median method is 0.5 since all the partial

median computations are performed over the entire data set. Computation of the median is  $O(n \log n)$  for practical purposes, and thus the time complexity of the RM method is high, of  $O(n^p \log^p n)$  order. The Gaussian efficiency of the repeated median method was found experimentally as being only around 0.6 (Siegel 1982). The repeated median is not affine equivariant, that is, a linear transformation of the explanatory variables does affect the estimate (Rousseeuw & Leroy 1987, p. 152).

The least median of squares (LMedS) robust regression method proposed by Rousseeuw (1984) also achieves 0.5 breakdown point. (We prefer to use the LMedS notation instead of LMS, which is used in the statistical literature, to avoid confusion with the usage in the image processing field where LMS stands for least mean squares.) The relative efficiency of the LMedS method can be improved by combining it with least-squares based techniques. The time complexity can be reduced by a Monte Carlo type speed-up technique. We will return to this estimator in section 3.

#### 2.1 Robust Methods in Computer Vision

Without trying to give an exhaustive survey, in this section we enumerate some of the computer vision applications using the robust techniques mentioned above. For the most recent results see the *Proceedings of the International Workshop on Robust Computer Vision*, Seattle, WA, October 1990.

Median- and trimmed-mean based local operators (Lestimators) have been employed in computer vision for a long time. See for example (Bovik et al. 1987) and (Coyle et al. 1989) for literature reviews on gray-level image smoothing and noise removal.

In the median of intercepts line fitting method of Kamgar-Parsi et al. (1989), similar to (Theil 1950) pairwise median, the intercept and slope is computed for every pair of points and the medians of the resulting lists are the estimates of the two parameters.

Recently M-estimators have also become popular in computer vision. Kashyap and Eom (1988) treated an image as a causal autoregressive model driven by a noise process assumed to be Gaussian with a small percent of the samples (at most 8%) contaminated by impulse noise, that is, outliers. The same technique was used by Koivo and Kim (1989) for classification of surface defects on wood boards. By employing M-estimators the parameters of the autoregressive process were iteratively refined simultaneously with cleaning the outliers in the noisy image. Besl et al. (1988) proposed a hierarchical scheme in which local fits of increasing degrees were obtained by M-estimators. The different fits were compared through a robust-fit quality measure to determine the optimal parameters. Haralick and Joo (1988) applied M-estimators to solve the correspondence problem between two sets of 2D perspective projections of model points in 3D. The correct-pose solution was then obtained with up to 30% of the pairs mismatched. A similar algorithm was used by Lee et al. (1989) for estimating 3D motion parameters.

The least median of squares estimator has also been used to solve computer vision problems. The high breakdown point makes it an optimum nonlinear interpolator (Kim et al. 1989 and Tirumalai & Schunck 1988), and it can be employed for image structure analysis in the piecewise polynomial (facet) domain [Meer et al. 1990]. Kumar and Hanson (1989) used least median of squares to solve the pose estimation problem. A variant of LMedS, the minimum-volume ellipsoid estimator, was used to develop a robust feature-space clustering technique which was then applied to histogram decomposition, Hough space analysis, and range image segmentation (Jolion et al. 1990 a, b).

#### 3 The Least-Median-of-Squares Method

Rousseeuw (1984) proposed the least-median-of-squares (LMedS) method in which the parameters are estimated by solving the nonlinear minimization problem.

min med 
$$r_i^2$$
 (7)

That is, the estimates must yield the smallest value for the median of squared residuals computed for the entire data set. An excellent application-oriented reference on the least-median-of-squares technique is the book of Rousseeuw and Leroy (1987) to which the reader is referred to for more details on the algorithm described below. The application of LMedS estimators to line fitting (p = 2) was studied in depth by Steele and Steiger (1986) and by Edelsbrunner and Souvaine (1988). They investigated the number of local minima of (7), and gave optimum time-complexity,  $O(n^2)$ , algorithms.

The LMedS minimization problem (7) cannot be reduced to a least-squares based solution, unlike the M-estimators (2). The least-median-of-squares minimization must be solved by a search in the space of possible estimates generated from the data. Since this space is too large, only a randomly chosen subset of points can be analyzed. We will return to this problem later in the section.

Assume that our data consist of *n* points from which we want to estimate the *p* regression coefficients  $\beta_j$ ,  $j = 0, 1, \ldots, (p - 1)$  of the linear model

$$z_i = \sum_{j=0}^{p-1} \beta_j x_i(i) \qquad i = 1, 2, \dots, n \quad (8)$$

The explanatory variables  $x_j(i)$  are monomials of different degrees in the sampling lattice coordinates. The model (8) thus represents polynomial surface fitting to the data.

Let a distinct *p*-tuple of data points be denoted by the indexes  $i_1, \ldots, i_p$ . For this *p*-tuple the values of p-1 parameters, all except the intercept,  $\hat{\beta}_0(i_1, \ldots, i_p)$ , can be computed by solving the *p* linear equations in  $\hat{\beta}_i(i_1, \ldots, i_p)$ ,  $j = 1, \ldots, (p-1)$ ,

$$z_i = \sum_{j=0}^{p-1} \hat{\beta}_j(i_1, \ldots, i_p) x_j(i) \quad i = i_1, \ldots, i_p \quad (9)$$

The intercept value  $\hat{\beta}_0(i_1, \ldots, i_p)$  is then found by solving the minimization problem

min med 
$$r_i^2$$
 given  $\hat{\beta}_j(i_1, \ldots, i_p), j = 1, \ldots, (p-1)$ 
(10)

The reduction of a multidimensional regression problem to one dimension is known as the *projection pursuit* technique, and has numerous applications in statistics (Efron 1988). Rousseeuw and Leroy (1987) have discussed the connection between this technique and the LMedS estimators. The goal of LMedS method is to identify the "outliers" in the data, that is, points severely deviating from the model. If these points can be discriminated in the projection along the direction of the intercept, the above procedure is optimum. We discuss the problem in more detail in section 4.

To solve (10) we must use a mode-estimation technique. The mode of a continuous probability distribution is the location of its maximum. In the case of a discrete, ordered sequence, the mode corresponds to the center of the subinterval having the highest density. Mode-seeking algorithms are well-known, see for example Press et al. (1988, p. 462). It can be shown that if the width of the search window is half the data size (i.e., |n/2|) the mode minimizes the median of the squared residuals (Rousseeuw & Leroy 1987, p. 169). Thus, if we define the projected sequence

$$s_i = z_i - \sum_{j=1}^{p-1} \hat{\beta}_j(i_1, \ldots, i_p) x_j(i)$$
 (11)

compute its mode, and take it as  $\hat{\beta}_0(i_1, \ldots, i_p)$ , we obtain the solution of (10).

The mode-seeking procedure also returns the width of the half-data size search window corresponding to the mode location. For *n* data points C(n, p), such window sizes are obtained and the smallest one yields the final LMedS estimate for the *p* regression coefficients of the model, that is, the solution of (7).

The breakdown point of the least median squares method is 0.5 because all the median computations are over the whole data set. The time-complexity of the method, however, is very high. There are  $O(n^p)$  ptuples and for each of them the sorting takes  $O(n \log n)$  time. Thus the amount of computation required for the basic LMedS algorithm is  $O(n^{p+1} \log n)$ , prohibitively large. Notice that this complexity is valid only if  $p \ge 2$ , since for p = 1 only sorting is required.

The time complexity is reduced to practical values when a Monte Carlo type speed-up technique is employed in which a  $Q \ll 1$  probability of error is tolerated. Let  $\epsilon$  be the fraction of data contaminated by outliers. Then the probability that all *m* different *p*tuples chosen at random will contain at least one or more outliers is

$$P = [1 - (1 - \epsilon)^{p}]^{m}$$
(12)

Note that 1 - P is the probability that at least one ptuple from the chosen m has only uncorrupted samples and thus the correct parameter values can be recovered. The smallest acceptable value for m is the solution of the equation P = Q, rounded upward to the closest integer, and is independent of n, the size of the data. The amount of computation becomes  $O(mn \log n)$ . This time-complexity reduction is very significant. For example, if p = 3, Q = 0.01, and  $\epsilon = 0.3$ , then m =11 for any n. Thus, when at most 30 percent of the data is contaminated by outliers, by choosing 11 triplets for the computation of the LMedS planar surface fit, the probability of having the whole set of triplets corrupted is 0.01. Rousseeuw and Leroy (1987) recommended taking a larger set of *p*-tuples than the number obtained from the probabilistic considerations. During the random sampling, identical p-tuples can be avoided without an increase in complexity (Mintz & Amir 1989). Additional speed-up relative to the original LMedS algorithm can be obtained by decomposition in both the spatial and parameter domains (Mintz et al. 1990).

When Gaussian noise is present in addition to outliers, the relative efficiency of the LMedS method is low. Rousseeuw (1984) has shown that the LMedS method converges for large sample sizes as  $n^{-1/3}$ , much more slowly than the usual  $n^{-1/2}$  for maximum likelihood estimators. To compensate for this deficiency he proposed combining the LMedS method with a weighted least squares procedure which has high Gaussian efficiency. Either one-step weighted least squares or an M-estimator with Hampel's redescending function can be employed but the latter appears to be less effective (Rousseeuw & Leroy 1987). Simultaneous presence of significant Gaussian noise and numerous outliers decreases the reliability of the LMedS estimates as is discussed in section 4.

The robust standard deviation estimate

$$\hat{\sigma} = 1.4826 \left(1 + \frac{5}{n-p}\right) \mod_{i} \sqrt{r_i^2}$$
(13)

can be immediately obtained since the median of the residual is the value returned by the LMedS procedure for the final parameter estimates. The factor 1.4826 is for consistent estimation in the presence of Gaussian noise, and the term 5/(n - p) is recommended by Rousseeuw and Leroy (1987) as a finite sample correction.

Based on the robust LMedS model and the standard deviation estimate binary weights can be allocated to the data points:

$$w_i = \begin{cases} 1 & \frac{|r_i|}{\hat{\sigma}} \le 2.5 \\ 0 & \frac{|r_i|}{\hat{\sigma}} > 2.5 \end{cases}$$
(14)

The data points having  $w_i = 1$  are inliers, that is, they belong to the assumed model. Points having  $w_i = 0$ are outliers and should not be further taken into consideration. The weighted least square estimates are then the solutions of the minimization problem

$$\min \sum_{i} w_{i} r_{i}^{2}$$
(15)

and are obtained by using any available programming package.

The presence of noise corrupting all the samples, however, may create severe artifacts. Since this is the case in numerous computer vision applications, in the next section we discuss the problem in detail.

# 4 Analysis of the LMedS Technique: Comparison with RANSAC

The LMedS estimates are computed by minimizing the cost function (7) through a search in the space of possible solutions. Subsets of the data are chosen by random sampling and for each subset a model is estimated. The number of points in the subset is equal to the number of unknown model parameters, thus yielding closed form solutions. If the number of points used is increased, in which case the model must be estimated by least squares, no improvement in the final LMedS estimates can be expected (Meer et al. 1990).

The search technique employed to find the LMedS estimates (projection pursuit) is common for several high breakdown point robust estimators (Rousseeuw & Leroy 1987). A similar processing principle was proposed independently by Fischler and Bolles (1981) for solving computer vision problems in the RANSAC paradigm. They too compute a model by solving a system of equations defined for a randomly chosen subset of points. All the data is then classified relative to this model. The points within some error tolerance are called the consensus set of the model. If the cardinality of the consensus set exceeds a threshold, the model is accepted and its parameters recomputed based on the whole consensus set. If the model is not accepted a new set of points is chosen and the resulting model is tested for validity. The error tolerance and the consensus set acceptance threshold must be set a priori.

In LMedS, the computed model is allocated an error measure, the median of the squared residuals. Several models are tried and the one yielding the minimum error is retained. The points are classified into inliers and outliers only relative to this final fit. The LMedS model can then be refined by a least squares procedure on the inlier data points. By its definition the LMedS estimator will always return at least 50 percent of the data points as inliers. The error tolerance is set automatically by the estimated robust standard deviation.

The LMedS estimator has a 0.5 breakdown point. Without a priori information about the structure of the data, in RANSAC the size of an acceptable consensus set also cannot be less than half the data size. For example, suppose that a constant must be estimated from data having only two values. Then, if the cardinality threshold on the consensus set is less than half the data size, the RANSAC procedure may validate either of the two constants.

To recover a model representing the relative and not absolute majority in the data, additional procedures should be implemented. Bolles and Fischler (1981) proposed the use of nonparametric tests for detecting the "white" structure of residuals. A decomposition technique applied in both the spatial and parameter domains (Mintz et al. 1990) also succeeds to return reliable estimates when less than half the data carry the model.

The maximum number of subsets required to validate a model for a tolerated probability of error is established by the same relation (12) in both LMedS and RANSAC. In LMedS, the algorithm will process that many *p*-tuples; in RANSAC, consensus may be found earlier. However, if a tolerated residual error threshold is used in LMedS, a stopping criterion similar to that of RANSAC is obtained. Using the statistical terminology, RANSAC maximizes the number of inliers while seeking the best model, and LMedS minimizes a robust error measure of that model. The two criteria are not mathematically equivalent, but since in the general case at least half of the data should become inliers, they yield very similar results.

We conclude that LMedS and RANSAC, in spite of being independently developed in different research areas, are based on similar concepts. The only difference of significance is that the LMedS technique generates the error measure during the estimation procedure, while RANSAC must be supplied with it. This is an important feature when the noise is not homogeneous, that is, the model is selectively corrupted. The robustness of both LMedS and RANSAC in the presence of severe deviations from the sought model is contingent upon the existence of at least one subset of data carrying the correct model. When noise corrupts all the data (e.g., Gaussian noise) the quality of initial model estimates degrades and could lead to incorrect decisions. The problem is emphasized when the model contains more parameters to represent the data than is necessary. That is, the order of the model is incorrect.

Consider again the two-valued data case, specifically, a step edge of amplitude h, the difference between the two values. Assume the data is evenly split, that is,  $\epsilon$  is close to 0.5. All the data is corrupted by a significant,

zero-mean noise process having values with significant probability in (-a, a), with a close to h/2. Let the RANSAC error threshold be a, and recall that this value must be chosen a priori. A first-order model is employed (p = 3) and thus planar fits are sought. The noise makes the models computed from triplets of data points unreliable. A model representing a tilted plane (similar to what the nonrobust least squares procedure would recover from all the data) already yields an absolute majority of points within the error tolerance, and is accepted by RANSAC. Reducing the error threshold decreases the expected number of points within the bounds, and thus a satisfactory consensus set size may not be obtained for the correct model. With high probability RANSAC will return an incorrect decision.

The LMedS procedure will also err. Most of the squared residuals relative to a tilted plane are between  $(0, a^2)$  and thus the median of the squared residuals is much less than  $a^2$ . Relative to the correct, horizontal plane slightly more than half of the squared residuals are between  $(0, a^2)$  with the median being close to  $a^2$ . Thus the LMedS algorithm too will prefer the tilted plane to the correct solution. For experimental data on this artifact of the LMedS estimators, and how it can be eliminated by a two-stage procedure, see Meer et al. (1990) and Mintz et al. (1990).

The robustness of the RANSAC paradigm, and the high breakdown point of the LMedS estimators, are thus meaningful only in situations when at least half of the data lie close to the desired model. For LMedS, in this case projection into the  $\beta_0$  direction results in a well-defined mode and there is no need for a complete projection pursuit procedure in which the optimum projection direction is also sought.

When LMedS local operators are applied to model discontinuities the number of tolerated outliers decreases. Consider again the noiseless, ideal step edge to which a  $5 \times 5$  robust local operator is applied. Assume that 10 pixels in the window belong to the edge (high amplitude) and 15 to the background (low amplitude) and that the center of the window falls on a background pixel. The operator returns the value of the majority of pixels, that is, the low amplitude of the background.

The image is then corrupted with fraction  $\epsilon = 0.2$ of asymmetric noise driving the corrupted samples into saturation at the upper bound. Without loss of generality we can assume that only 3 of the pixels belonging to the background were corrupted in the processing window. There are now 13 pixels with high amplitudes and the operator returns, incorrectly, a high value similar to the amplitude of the edge. Thus, even when the fraction of corrupted points is much smaller than the theoretical breakdown point of a robust estimator, the operator may systematically fail near transitions between homogeneous regions in images. At transitions, samples of one region are outliers (noise) when fitting a model to the other region, and a small fraction of additional noise may reverse the class having the majority. The size of the local operator also limits  $\epsilon$ , the maximum amount of tolerated contamination, but this artifact has less importance for images where the window operators have relative large supports.

It should not be concluded from this section that high breakdown-point estimation procedures are not useful for computer vision. We have emphasized their sensitivity to *piecewise models*, that is, to data containing discontinuities, in order to increase awareness of possible pitfalls. In section 2.1 several successful applications of the LMedS estimators were mentioned. Recently we have proposed a new technique for noisyimage analysis in the piecewise polynomial domain combining simple nonrobust and robust processing modules. The method has a high breakdown point but avoids the problems discussed above. In the next section we use a simple application of the LMedS estimators, window-operator-based smoothing, to compare its performance with M-estimators.

# 5 Image Smoothing: LMedS vs. M-estimators

Signals unchanged by the application of an operator are called root signals of that operator. The existence of root signals assures the convergence of iterative nonlinear filtering procedures (Fitch et al. 1985). Their importance is also recognized in computer vision (e.g., Haralick & Watson 1981; Owens et al. 1989). In one dimension, a noiseless piecewise polynomial signal is a root signal of an LMedS-based window operator of degree equal to the highest polynomial degree. This property is in fact true for any 0.5 breakdown point estimator. Indeed, since the input is an ideal piecewise polynomial signal, a contiguous group of  $\lceil n/2 \rceil$  pixels always carries information about the same polynomial. The 0.5 breakdown point assures that the estimated regression coefficients are those of this polynomial. The value returned by the operator is the

value of the polynomial at the window center and is identical with the input. The noiseless one-dimensional piecewise polynomial signal remains undistorted. This root-signal property is not necessarily valid in two dimensions. When the window is centered on a corner, it is not always the case that 51% of the pixels belong to the surface in the window center. Before proceeding to compare root-signal properties of the Mestimators and the LMedS estimator we give a short description of the former.

M-estimators have already been introduced in section 2. The minimization problem (2) is reduced to reweighted least squares by differentiating (2) with respect to the sought regression coefficients. After some simple manipulations, the following expression for the weights is obtained:

$$w_i = \frac{1}{r_i} \frac{\mathrm{d}\rho(r_i)}{\mathrm{d}r_i} \tag{16}$$

where  $r_i$  is the residual of the *i*th data point. Since the  $\rho$  functions have a minimum at  $r_i = 0$ , expression (16) can always yield  $w_i = 1$  at the origin. All the weights are positive and between 0 and 1. Different  $\rho$  functions result in different weights. We consider here two such functions.

The first robust M-estimator was proposed by Huber in 1964 (see Huber 1981). It is known as the minimax M-estimator and has least squares behavior for small residuals, and the more robust least-absolute-values behavior for large residuals. The change in behavior is controlled by a tuning constant  $c_H$  and the locally estimated noise standard deviation

$$\hat{\sigma} = 1.4826 \mod |r_i - \mod r_i| \tag{17}$$

where med denotes the median taken over the entire window, and the factor 1.4826 compensates for the bias of the median estimator in Gaussian noise. Note that (17) is a robust estimator of  $\sigma$  and can tolerate up to half the data being corrupted. The minimax weight function has the expression

$$w_{H,i} = \begin{cases} 1 & |r_i| \le c_H \hat{\sigma} \\ \frac{c_H \hat{\sigma}}{|r_i|} & |r_i| > c_H \hat{\sigma} \end{cases}$$
(18)

and is shown in figure 1b together with its  $\rho$  function (figure 1a). Note that the constant weight corresponds to the central (least squares) region, and in the least-

absolute-values region the large residuals have hyperbolically decreasing weights. To achieve 95% asymptotic efficiency for Gaussian noise Holland and Welsch (1977) recommended  $c_H = 1.345$ .

In the class of redescending M-estimators, the large residuals have zero weights thus improving the outlier rejection properties. The biweight

$$w_{B,i} = \begin{cases} \left[ 1 - \left(\frac{r_i}{c_B \hat{\sigma}}\right)^2 \right]^2 & |r_i| \le c_B \hat{\sigma} \\ 0 & |r_i| > c_B \hat{\sigma} \end{cases}$$
(19)

proposed by Beaton and Tukey (1974) is a well-known example of this class. The weight function is shown in figure 2b and the  $\rho$  function it was derived from in figure 2a. Holland and Welsch (1977) recommended  $c_B$ = 4.685 to assure superior performance for Gaussian noise. This tuning constant value was obtained from a Monte Carlo study employing homogeneous data (i.e., from one model). In computer vision we are often dealing with piecewise models (i.e., data with discontinuities) and smaller tuning constants should be used to discriminate the transitions. In our experiments we took  $c_B = 2$ .

The weights cannot be computed without a standard deviation estimate, which in turn requires the estimation of an initial fit. The quality of the initial fit is of paramount importance, especially for redescending Mestimators. The use of robust estimators like least absolute deviations or LMedS is recommended (Hampel et al., 1986). We are interested in the outlier rejection capability of the M-estimators, since this property assures the undistorted recovery of the input. Therefore to facilitate the comparison between M-estimators and LMedS we employ an unweighted least squares procedure (all  $w_i = 1$ ) to obtain the initial M-estimates. The residual relative to this fit can be computed and the standard deviation estimate obtained. Each data point is then allocated a weight and in the next iteration the new set of parameters is obtained by weighted least squares. The standard deviation estimate  $\hat{\sigma}$  should not be updated during the iterations (Holland & Welsch 1977). At consecutive iterations, samples yielding large residuals (outliers) have their weights reduced, and the estimates are computed mostly from values distributed around the true surface. Often for the latter samples a Gaussian distribution is assumed.



Fig. 1. The Huber minimax M-estimator. (a) The  $\rho$  function. (b) The weight function.



Fig. 2. The biweight M-estimator. (a) The  $\rho$  function. (b) The weight function.

The following quality measure was used by us as a stopping criterion for the iterations:

$$E^{(1)} = \sqrt{\frac{\sum_{i=1}^{n} w_{i} r_{i}^{2}}{\sum_{i=1}^{n} w_{i}}}$$
(20)

where the index l denotes the current iteration. Convergence is achieved if

$$|E^{(l)} - E^{(l-1)}| < 0.001 \qquad l = 1, 2, \dots$$
 (21)

The number of allowed iterations was bounded to 25.

A  $100 \times 100$  synthetic image containing several polyhedral objects was used as an example. Only planar surfaces are present, joined at either step or roof edges. In figure 3a the gray level and in figure 3b the wire-frame representation of the image is shown. The image was smoothed with  $5 \times 5$  window operators by replacing the value of the center pixel with the intercept of the estimated plane. Note that as an artifact of this simple smoothing procedure a systematic error is introduced at the corners where the center pixel does not belong to the surface containing the majority.

We have used least squares (figure 4a), the minimax M-estimator (figure 4b), the biweight M-estimator (figure 4c), and the LMedS estimator (figure 4d) in the

window. The M-estimators did not converge for less than one percent of the pixels. For each method the smoothing error, defined as square root of the sum of the squared differences between the input and output values of a pixel, was computed. (See table 1.)

As is well known, the least squares fit fails at discontinuities and significant blurring of the input occurs. This fit is also the initial estimate for the M-estimators. It yields an incorrect, high standard deviation estimate and the iterative procedure cannot completely recover the input. The two M-estimators also blur edges, although in a lesser amount. The minimax M-estimator has nonzero weights for all the pixels and thus introduces slightly more blurring at edges than the biweight estimator. All the errors made by the LMedS estimator are due to the artifact at the corners. This artifact can be eliminated by cooperative processes (Meer et al. 1990; Mintz et al. 1990).

Thus due to its high breakdown point only the LMedS window operator succeeds in preserving a piecewise polynomial image. This property is useful when the operator is used as a nonlinear interpolator. We have compared in this section the raw performances of the estimators. Their properties can be improved when using them as building blocks in more complex algorithms. The smoothing performance of the Mestimators is corrected when several weight functions are employed at successive iterations and the model order is sequentially increased under the control of a



Fig. 3. A synthetic 100×100 image. (a) Gray-level representation. (b) Wireframe representation.

quality measure (Besl et al. 1988). Similarly, better tolerance of the LMedS smoothing operators to noise corrupting all the samples can be achieved (Mintz et al. 1990).

Table 1. Root mean square smoothing errors.

Method	Least Squares	Minimax	Biweight	LMedS
Error	944	922	855	392

## **6** Conclusion

We have presented a review of the different robust regression methods with special emphasis on the least median of squares estimators. We did not present applications since most of them involve technical details beyond the goal of this article. A literature survey of computer vision algorithms employing robust techniques was given in section 2.1.

The desirable high breakdown point of the LMedS estimators is contingent upon the sought model being represented by weakly corrupted data. Using a probabilistic speed-up technique, the computation of LMedS estimates is feasible, although more demanding than that of M-estimates. The latter, however, requires a reliable initial estimate, possibly the output of an LMedS algorithm. We conclude that application of robust estimators to real computer vision problems, with noisy data not perfectly representing the assumed model, should be preceded by a careful analysis of the problem. If a clear dichotomy exists between the "good" and "bad" data (as is usually the case in statistics) methods like LMedS are extremely useful. If good and bad data cannot be discriminated, additional precautions should be taken.

# Acknowledgments

We would like to thank the anonymous reviewers whose observations substantially improved the presentation of the paper. The support of the Defense Advanced Research Projects Agency and U.S. Army Night Vision and Electro Optics Laboratory under Grant DAAB07-86-K-F073, and of the National Science Foundation under Grant DCR-86-03723, is gratefully acknowledged.

# References

- Andrews, D.F. 1974. A robust method for multiple linear regression. *Technometrics* 16:523-531.
- Beaton, A.E., and Tukey, J.W. 1974. The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data. *Technometrics* 16:147–185.



Fig. 4. Window smoothing of the image in figure 3. (a) Least squares operator. (b) Minimax M-estimator. (c) Biweight M-estimator. (d) LMedS estimator.

- Besl, P.J., Birch, J.B., and Watson, L.T. 1988. Robust window operators. Proc. 2nd Intern. Conf. Comput. Vision, Tampa, FL, pp. 591-600. See also, Mach. Vision Appl. 2:179-214, 1989.
- Bolles, R.C., and Fischler, M.A. 1981. A RANSAC-based approach to model fitting and its application to finding cylinders in range data. *Proc. 7th Intern. Joint Conf. Artif. Intell.*, Vancouver, Canada, pp. 637-643.
- Bovik, A.C., Huang, T.S., and Munson, Jr., D.C. 1987. The effect of median filtering on edge estimation and detection. *IEEE Trans. Patt. Anal. Mach. Intell.* PAMI-9:181-194.
- Brown, G.W., and Mood, A.M. 1951. On median tests for linear hypotheses. *Proc. 2nd Berkeley Symp. Math. Stat. Prob. J. Neyman* (ed.), University of California Press: Berkeley and Los Angeles, pp. 159-166.
- Cheng, K.S., and Hettmansperger, T.P. 1983. Weighted least-squares rank estimates. *Commun. Stat.* A12:1069-1086.
- Coyle, E.J., Lin, J.H., and Gabbouj, M. 1989. Optimal stack filtering and the estimation and structural approaches to image processing. *IEEE Trans. Acoust. Speech, Sig. Process.* 37:2037– 2066.

- Edelsbrunner, H., and Souvaine, D.L. 1988. Computing median-ofsquares regression lines and guided topological sweep. Report UIUCDCS-R-88-1483. Department of Computer Science, University of Illinois at Urbana-Champaign.
- Efron, B. 1988. Computer-intensive methods in statistical regression. SIAM Review 30:421-449.
- Fischler, M.A., and Bolles, R.C. 1981. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Commu. ACM* 24:381–395.
- Fitch, J.P., Coyle, E.J., and Gallagher, Jr., N.C. 1985. Root properties and convergence rates of median filters," *IEEE Trans. Acoust.*, *Speech, Sig. Process.* 33:230-240.
- Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J., and Stahel, W.A. 1986. Robust Statistics: An Approach Based on Influence Functions. Wiley: New York.
- Haralick, R.M., and Watson, L. 1981. A facet model for image data. Comput. Graph. Image Process. 15:113-129.
- Haralick, R.M., and Joo, H. 1988. 2D-3D pose estimation. Proc. 9th Intern. Conf. Patt. Recog. Rome, pp. 385-391. See also, IEEE Trans. Systems, Man, Cybern. 19:1426-1446, 1989.
- Heiler, S. 1981. Robust estimates in linear regression—A simulation approach. In *Computational Statistics*. H. Büning and P. Naeve (eds.), De Gruyter, Berlin, pp. 115–136.
- Holland, P.W., and Welsch, R.E. 1977. Robust regression using iteratively reweighted least squares. *Commun. Stat.* A6:813-828.
- Huber, P.J. 1981. Robust Statistics. Wiley: New York.
- Jaeckel, L.A. 1972. Estimating regression coefficients by minimizing the dispersion of residuals. *Ann. Math. Stat.* 43:1449-1458.
- Johnstone, I.M., and Velleman, P.F. 1985. The resistant line and related regression methods. J. Amer. Stat. Assoc. 80:1041-1059.
- Jolion, J.M., Meer, P., and Rosenfeld, A. 1990a. Generalized minimum volume ellipsoid clustering with applications in computer vision. *Proc. Intern. Workshop Robust Comput. Vision*, Seattle, WA, pp. 339-351.
- Jolion, J.M., Meer, P., and Bataouche, S., 1990b. Range image segmentation by robust clustering. CAR-TR-500. Computer Vision Laboratory, University of Maryland, College Park.
- Kamgar-Parsi, B., Kamgar-Parsi, B., and Netanyahu, N.S. 1989. A nonparametric method for fitting a straight line to a noisy image. *IEEE Trans. Patt. Anal. Mach. Intell.* PAMI-II:998-1001.
- Kashyap, R.L., and Eom, K.B. 1988. Robust image models and their applications. In Advances in Electronics and Electron Physics, vol. 70, Academic Press: San Diego, CA, pp. 79–157.
- Kim, D.Y., Kim, J.J., Meer, P., Mintz, D., and Rosenfeld, A. 1989. Robust computer vision: A least median of squares based approach. *Proc. DARPA Image Understanding Workshop*, Palo Alto, CA, pp. 1117-1134.

- Koivo, A.J., and Kim, C.W. 1989. Robust image modeling for classification of surface defects on wood boards. *IEEE Trans. Syst. Man, Cybern.* 19:1659–1666.
- Kumar, R., and Hanson, A.R. 1989. Robust estimation of camera location and orientation from noisy data having outliers. *Proc. Workshop on Interpretation of 3D Scenes*, Austin, TX, pp. 52-60.
- Lee, C.N., Haralick, R.M., and Zhuang, X. 1989. Recovering 3-D motion parameters from image sequences with gross errors. *Proc. Workshop on Visual Motion*, Irvine, CA, pp. 46–53.
- Li, G. 1985. Robust regression. In D.C. Hoaglin, F. Mosteller, and J.W. Tukey (eds.), *Exploring Data Tables, Trends and Shapes*. Wiley: New York, pp. 281-343.
- Meer, P., Mintz, D., and Rosenfeld, A. 1990. Least median of squares based robust analysis of image structure. *Proc. DARPA Image Understanding Workshop*, Pittsburgh, PA, pp. 231-254.
- Mintz, D., and Amir, A. 1989. Generalized random sampling. CAR-TR-441. Computer Vision Laboratory, University of Maryland, College Park.
- Mintz, D., Meer, P., and Rosenfeld, A. 1990. A fast, high breakdown point robust estimator for computer vision applications. *Proc. DARPA Image Understanding Workshop*, Pittsburgh, PA, pp. 255-258.
- Mosteller, F., and Tukey, J.W. 1977. Data Analysis and Regression. Addison-Wesley: Reading, MA.
- Owens, R., Venkatesh, S., and Ross, J. 1989. Edge detection is a projection. Patt. Recog. Lett. 9:233-244.
- Press, W.H., Flannery, B.P., Teukolsky, S.A., and Vetterling, W.T. 1988. Numerical Recipes. Cambridge University Press, Cambridge, England.
- Rousseeuw, P.J. 1984. Least median of squares regression. J. Amer. Stat. Assoc. 79:871-880.
- Rousseeuw, P.J., and Leroy A.M. 1987. Robust Regression & Outlier Detection. Wiley: New York.
- Siegel, A.F. 1982. Robust regression using repeated medians. Biometrika 69:242-244.
- Steele, J.M., and Steiger, W.L. 1986. Algorithms and complexity for least median of squares regression. *Discrete Appl. Math.* 14:93–100.
- Theil, H. 1950. A rank-invariant method of linear and polynomial regression analysis (parts 1-3). Ned. okad. Wetensch. Proc. ser. A 53:386-392, 521-525, 1397-1412.
- Tirumalai, A., and Schunck, B.G. 1988. Robust surface approximation using least median squares regression. CSE-TR-13-89. Artificial Intelligence Laboratory, University of Michigan, Ann Arbor.