Chapter 6

MRF Model with Robust Statistics

Robust statistical methods (Tukey 1977; Huber 1981; Rousseeuw 1984) are tools for statistics problems in which *outliers* are an issue. It is well known that the least squares (LS) error estimates can be arbitrarily wrong when outliers are present in the data. A robust procedure is aimed at making solutions insensitive to the influence of outliers. That is, its performance should be good with all-inlier data and should deteriorate gracefully with increasing number of outliers. The mechanism by which robust estimators deal with outliers is similar to that of the discontinuity adaptive MRF prior model studied in the previous chapter. This chapter provides a comparative study (Li 1995a) of the two kinds of models based on the results from the DA model and presents an algorithm (Li 1996b) to improve the stability of the robust M-estimator to the initialization.

The conceptual and mathematical comparison comes naturally from the parallelism of the two models: Outliers cause a violation of a distributional assumption, while discontinuities cause a violation of the smoothness assumption. Robustness to outliers is in parallel to adaptation to discontinuities. Detecting outliers corresponds to inserting discontinuities. The similarity of the two models suggests that results in either model could be used for the other.

Probably for this reason,there have seen considerable interests in applying robust techniques to solving image and vision problems. Kashyap and Eom (1988) developed a robust algorithm for estimating parameters in an autoregressive image model where the noise is assumed to be a mixture of a Gaussian and an outlier process. Shulman and Herve (1989) proposed to use Huber's robust M-estimator (Huber 1981) to compute optical flow involving discontinuities. Stevenson and Delp (1990) used the same estimator for curve fitting. Besl, Birch, and Watson (1988) proposed a robust M window

S.Z. Li, *Markov Random Field Modeling in Image Analysis*,
Advances in Pattern Recognition, DOI: 10.1007/978-1-84800-279-1_6,

operator to prevent smoothing across discontinuities. Haralick, Joo, Lee, Zhuang, Vaidya, and Kim (1989), Kumar and Hanson (1989) and Zhuang, Wang, and Zhang (1992) used robust estimators to find pose parameters. Jolion, Meer, and Bataouche (1991) used the robust minimum volume ellipsoid estimator to identify clusters in feature space. Boyer, Mirza, and Ganguly (1994) present a procedure for surface parameterization based on a robust M-estimator. Black and Anandan (1993) and Black and Rangarajan (1994) applied a robust operator not only to the smoothness term but also to the data term. Li (1995a) presented a comparative study on robust models and discontinuity adaptive MRF models. He also devised a method for stabilizing robust M-estimation w.r.t. the initialization and convergence (Li 1996b).

A robust location estimator, which essentially seeks the mode of an outlier-contaminated distribution, can be extended to perform data clustering. In this connection, the mean shift algorithm (Fukunaga 1990) has been used successfully in vision problems such as segmentation (Cheng 1995; Comaniciu and Meer 1997; Comaniciu and Meer 1999).

As is well known, robust estimation procedures have a serious problem in that the estimates are dependent on the initial estimate value; this problem has been overcome by applying the principle of the graduated nonconvexity (GNC) method (Blake and Zisserman 1987) for visual reconstruction. The exchange of theoretical results and practical algorithms is useful to the image and vision communities because both MRF and robust models have applications in the4 areas.

6.1 The DA Prior and Robust Statistics

What do we mean by discontinuities and outliers? Unfortunately, their definitions are usually ambiguous. What we are certain of is that the likelihood of a discontinuity between a pair of neighboring pixels is related to the difference in pixel labels (such as pixel values), and an outlier is related to the distance between the location of the datum and the estimated value. Where the label difference is very large, there is likely to be a discontinuity between the pixels, and where the datum is very far from the cluster, it is likely an outlier.

A more concrete comparison can be made by analyzing the adaptation (to discontinuities) and the robustness (to outliers) in mathematical terms (Li 1995a). The adaptation is realized as follows. The interaction between related (e.g, neighboring) points must be decreased as the violation of the relational bond between them is increased and prohibited in the limit. This is true of both the MRF and the robust models. We give the necessary condition for such adaptation and then, based on this condition, a definition of a class of adaptive interaction functions for both models. The definition captures the essence of the adaptation ability and is general enough to offer in theory infinitely many suitable choices of such functions.

The problem of discontinuities and outliers also exists in other areas. In model-based object recognition, for example, there are two related subproblems: first, separating the scene into different parts, each being due to a single object; and second, finding feature correspondences between each separate part of the scene and an object. The two subproblems have to be solved somewhat simultaneously. The process of matching while separating is similar to reconstruction with discontinuities and estimation with outliers. Indeed, matches to one object can be considered as outliers w.r.t. matches to a different object. Different groups of matches should not be constrained to each other. The separation can be done by inserting "discontinuities" between different groups. This view can be regarded as a generalization of the weak constraint (Hinton 1978; Blake and Zisserman 1987).

6.1.1 Robust M-Estimator

Robust statistical methods (Tukey 1977; Huber 1981) provide tools for statistics problems in which underlying assumptions are inexact. A robust procedure should be insensitive to departures from the underlying assumptions caused, for example, by *outliers*. That is, it should have good performance under the underlying assumptions, and the performance should deteriorate gracefully as the situation departs from the assumptions. Applications of robust methods in vision are seen in image restoration, smoothing and segmentation (Kashyap and Eom 1988; Jolion et al. 1991; Meer et al. 1991), surface and shape fitting (Besl et al. 1988; Stein and Werman 1992), and pose estimation (Haralick et al. 1989), where outliers are an issue.

There are several types of robust estimators. Among them are the M-estimator (maximum likelihood estimator), L-estimator (linear combinations of order statistics), R-estimator (estimator based on rank transformation) (Huber 1981), RM estimator (repeated median) (Siegel 1982) and LMS estimator (estimator using the least median of squares) (Rousseeuw 1984). We are concerned with the M-estimator.

The essential form of the M-estimation problem is the following. Given a set of m data samples $d = \{d_i \mid 1 \leq i \leq m\}$, where $d_i = f + \eta_i$, the problem is to estimate the location parameter f under noise η_i. The distribution of η_i is not assumed to be known exactly. The only underlying assumption is that $\eta_1, \ldots, \eta_m$ obey a symmetric, independent, identical distribution (symmetric i.i.d.). A robust estimator has to deal with departures from this assumption.

Let the residual errors be $\eta_i = d_i - f$ $(i = 1, \ldots, m)$ and the error penalty function be $g(\eta_i)$. The M-estimate f^* is defined as the minimum of a global error function

$$f^* = \arg\min_f E(f) \tag{6.1}$$

where

$$E(f) = \sum_i g(d_i - f) \tag{6.2}$$

Table 6.1: Robust functions.

Type	$h_\gamma(\xi)$	$g_\gamma(\xi)$	Range of ξ
Tukey	$= \begin{cases} (1-\xi^2)^2 \\ 0 \end{cases}$	$= \begin{cases} [1-\left(1-\xi^2\right)^3]/6 \\ 1/6 \end{cases}$	$\begin{array}{l} \lvert\xi\rvert \le 1 \\ \lvert\xi\rvert > 1 \end{array}$
Huber	$= \begin{cases} 1 \\ \tau\frac{\mathrm{sgn}(\xi)}{\xi} \end{cases}$	$= \begin{cases} \xi^2 \\ 2\tau\lvert\xi\rvert - \tau^2 \end{cases}$	$\begin{array}{l} \lvert\xi\rvert \le \tau \\ \lvert\xi\rvert > \tau \end{array}$
Andrews	$= \begin{cases} \frac{\sin(\pi\xi)}{\pi u} \\ 0 \end{cases}$	$= \begin{cases} [1-\cos(\pi\xi)]/\pi^2 \\ 1/\pi^2 \end{cases}$	$\begin{array}{l} \lvert\xi\rvert \le 1 \\ \lvert\xi\rvert > 1 \end{array}$
Hampel	$= \begin{cases} 1 \\ a\frac{\mathrm{sgn}(\xi)}{\xi} \\ a\frac{c-\lvert\xi\rvert}{c-b}\frac{\mathrm{sgn}(\xi)}{\xi} \\ 0 \end{cases}$	$= \begin{cases} u^2/2 \\ a\lvert u\rvert - a^2/2 \\ ab - a^2/2 + \\ (c-b)a/2\left[1-\left(\frac{c-\lvert\xi\rvert}{c-b}\right)\right] \\ ab - a^2/2 + (c-b)a/2 \end{cases}$	$\begin{array}{l} \lvert\xi\rvert \le a \\ a < \lvert\xi\rvert \le b \\ b < \lvert\xi\rvert \le c \\ \\ \lvert\xi\rvert > c \end{array}$

To minimize (6.2), it is necessary to solve the equation

$$\sum_i g'(d_i - f) = 0 \tag{6.3}$$

This is based on *gradient descent*. When $g(\eta_i)$ can also be expressed as a function of η_i^2, its first derivative can take the form

$$g'(\eta_i) = 2\eta_i h(\eta_i) = 2(d_i - f)h(d_i - f) \tag{6.4}$$

where $h(\eta)$ is an even function. In this case, the estimate f^* can be expressed as the weighted sum of the data samples

$$f^* = \frac{\sum_i h(\eta_i)\ d_i}{\sum_i h(\eta_i)} \tag{6.5}$$

where h acts as the weighting function. This algorithm can be derived by using half-quadratic (HQ) optimization to be presented in Section 6.1.6.

In the LS regression, all data points are weighted the same with $h_\gamma(\eta) = 1$ and the estimate is $f^* = \frac{1}{m}\sum_{i=1}^m d_i$. When outliers are weighted equally as inliers, it will cause considerable bias and deterioration in the quality of the estimate. In robust M-estimation, the function h provides adaptive weighting. The influence from d_i is decreased when $|\eta_i| = |d_i - f|$ is very large and suppressed when it is infinitely large.

Table 6.1 lists some robust functions used in practice where $\xi = \eta/\gamma$. They are closely related to the adaptive interaction function and adaptive potential function defined in (5.27) and (5.28). Figure 6.1 shows their qualitative shapes in comparison with the quadratic and the line process models (note that a

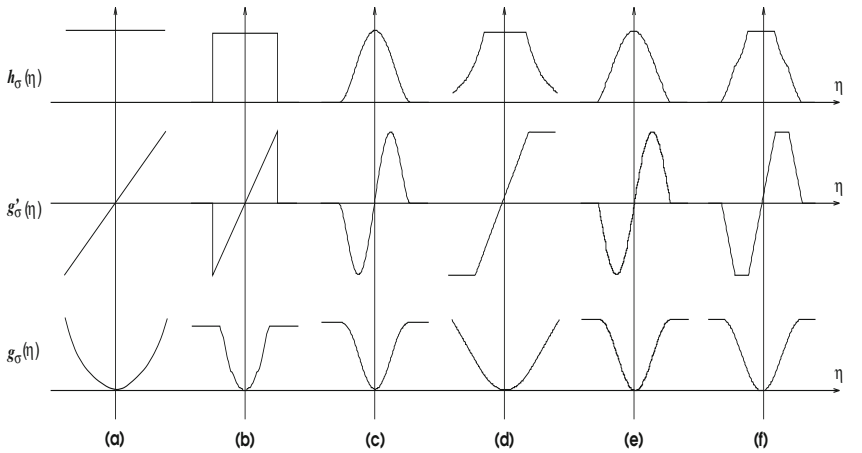

Figure 6.1: The qualitative shapes of potential functions in use. The quadratic prior (equivalent to LS) model in (a) is unable to deal with discontinuities (or outliers). The line process model (b), Tukey's (c), Huber's (d), Andrews' (e), and Hampel's (f) robust model are able to, owing to their property of $\lim_{\eta\to\infty} h_\gamma(\eta) = 0$. From (Li 1995a) with permission; ©1995 Elsevier.

trivial constant may be added to $g_\gamma(\eta)$). These robust functions are piecewise, as in the line process model. Moreover, the parameter γ in ξ is dependent on some scale estimate, such as the median of absolute deviation (MAD).

6.1.2 Problems with M-Estimator

Computationally, existing M-estimators have several problems affecting their performance. First, they are not robust to the initial estimate, a problem common to nonlinear regression procedures (Myers 1990), also encountered by vision researchers (Haralick et al. 1989; Meer et al. 1991; Zhuang et al. 1992). The convergence of the algorithm depends on the initialization. Even if the problem of convergence is avoided, the need for a good initial estimate cannot be ignored for convergence to the global estimate; this is because most M-estimators are defined as the global minimum of a generally *nonconvex* energy function and hence the commonly used gradient-based algorithms can get stuck at unfavorable local solutions. The M-estimator has the theoretical breakpoint of $\frac{1}{p+1}$, where p is the number of unknown parameters to be estimated, but, in practice, the breakpoint can be well below this value because of the problem of local minima.

Second, the definition of the M-estimator involves some scale estimate, such as the median of absolute deviation (MAD), and a parameter to be chosen. These are also sources of sensitivity and instability. For example, Tukey's biweight function (Tukey 1977) is defined as

$$h(\eta_i) = \begin{cases} \left(1 - \left(\frac{\eta_i}{cS}\right)^2\right)^2 & \text{if } |\eta_i| < cS \\ 0 & \text{otherwise} \end{cases} \tag{6.6}$$

where S is an estimate of the spread, c is a constant parameter, and cS is the scale estimate. Possible choices include $S = \text{median}\{\eta_i\}$ with c set to 6 or 9, and $S = \text{median}\{|\eta_i - \text{median}\{\eta_i\}|\}$ (median of absolute deviation (MAD)) with $c = 1.4826$ chosen for the best consistency with the Gaussian distribution. Classical scale estimates such as the median and MAD are not very robust. The design of the scale estimates is crucial and needs devoted study.

Furthermore, the convergence of the M-estimator often is not guaranteed. Divergence can occur when initialization or parameters are not chosen properly. Owing to the problems above, the theoretical breakdown point can hardly be achieved.

In the following, an improved robust M-estimator, referred to as the *annealing M-estimator* (AM-estimator), is presented to overcome the above problems (Li 1996b). It has two main ingredients: a redefinition of the M-estimator and a GNC-like annealing algorithm.

6.1.3 Redefinition of M-Estimator

Resemblances between M-estimation with outliers and adaptive smoothing with discontinuities have been noted by several authors (Besl et al. 1988; Shulman and Herve 1989; Black and Anandan 1993; Black and Rangarajan 1994; Li 1995a). We can compare the M-estimator with the DA model studied in Chapter 5. The influence of the datum d_i on the estimate f is proportional to $\eta_i h(\eta_i)$. This compares with the smoothing strength $f'h(f')$ given after (5.25). A very large $|\eta_i|$ value, due to d_i being far from f, suggests an outlier. This is similar to saying that a very large $|f'(x)|$ value is likely due to a step (discontinuity) in the signal there. The resemblance suggests that the definition of the DA model can also be used to define M-estimators (Li 1996b).

We replace the scale estimate in the M-estimator by a parameter $\gamma > 0$ and choose to use the adaptive interaction function h_γ and the adaptive potential function g_γ for the M-estimation. However, h_γ need only be C^0 continuous for the location estimation from discrete data. Theoretically, the definitions give an infinite number of suitable choices of the M-estimators. Table 5.1 and Fig. 5.1 showed four such possibilities. With h_γ and g_γ, we can define the energy under γ as

$$E_\gamma(f) = \sum_i g_\gamma(d_i - f) \tag{6.7}$$

and thereby the minimum energy estimate

$$f_\gamma^* = \arg\min_f E_\gamma(f) = \frac{\sum_i h_\gamma(\eta_i)\ d_i}{\sum_i h_\gamma(\eta_i)} \tag{6.8}$$

AM-Estimator
Begin Algorithm
set $t = 0$, $f_\gamma^{(1)} = f^{LS}$; choose initial γ;
do {
$t \leftarrow t + 1$;
compute errors $\eta_i = d_i - f_\gamma^{(t-1)}$, $\forall i$;
compute weighted sum $f_\gamma^{(t)} = \frac{\sum_i h_\gamma(\eta_i)\ d_i}{\sum_i h_\gamma(\eta_i)}$;
if ($|f_\gamma^{(t)} - f_\gamma^{(t-1)}| < \epsilon$) /* converged */
$\gamma \leftarrow$ lower(γ);
} until ($\gamma < \delta$) /* frozen */
$f^* \leftarrow f_\gamma^{(t)}$;
End Algorithm

Figure 6.2: The AM-estimation algorithm.

This defines a class of M-estimators that are able to deal with outliers as the traditional M-estimators do. Their performance in the solution quality is significantly enhanced by using an annealing procedure.

6.1.4 AM-Estimator

The annealing algorithm for the redefined M-estimator, called the AM-estimator (Li 1996b), is based on the idea of the GNC algorithm (Blake and Zisserman 1987). It aims to overcome the local minimum problem in the M-estimation (i.e., to obtain a good estimate regardless of the initialization). It also make the estimation free from parameters or at least insensitive to their choices.

The annealing is performed by continuation in γ, and the AM-estimator is defined in the limit

$$f^* = \lim_{\gamma \to 0^+} f_\gamma^* \tag{6.9}$$

An algorithm that implements the AM-estimator algorithm is given in Fig. 6.2. Initially, γ is set to a value high enough to guarantee that the corresponding APF $g_\gamma(\eta)$ is convex in an interval. With such a γ, it is easy to find the unique minimum of the global error function $E_\gamma(f)$ using the gradient-descent method, regardless of the initial value for f. The minimum is then used as the initial value for the next phase of minimization under a lower γ to obtain the next minimum. As γ is lowered, $g_\gamma(\eta)$ may no longer be convex and local minima may appear. However, if we track the global minima

for decreasing γ values, we may hopefully approximate the global minimum f^* under $\gamma \to 0^+$.

Obviously, whatever the initialization is, the first iteration always gives a value equal to the LS estimate

$$f^{LS} = \frac{1}{m} \sum_{i=1}^{m} d_i \tag{6.10}$$

This is because for $\gamma \to +\infty$, which guarantees the strict convexity, all weights are the same as $h_\gamma(\eta_i) = 1$. The initial γ is chosen to satisfy

$$|\eta_i| = |d_i - f^{LS}| < b_H(\gamma) \tag{6.11}$$

where $b_H(\gamma)$ $(= -b_L(\gamma))$ is the upper bound of the band in (5.29). This guarantees $g''_\gamma(\eta_i) > 0$ and hence the strict convexity of g_γ. The parameter γ is lowered according to the schedule specified by the function lower(γ). The parameters δ and ϵ in the convergence conditions are some small numbers. An alternative way is to decrease γ according to a fixed schedule regardless of whether or not $f^{(t)}$ converges at the current γ, which is equivalent to setting a big value for ϵ. In this case, the algorithm freezes after dozens of iterations. This quick annealing is used in our experiments.

The advantages of the AM-estimator are summarized below. First, the use of the annealing significantly improves the quality and stability of the estimate. The estimate is made independent of the initialization. Because the starting point for obtaining f^*_γ at current γ is the convergence point obtained with the previous γ value, the divergence problem with the traditional M-estimator is minimized. Second, the definition of the AM-estimator effectively eliminates scale parameters in the M-estimation because γ is finally set to zero (or a small number to whose value the final estimate is insensitive). This avoids the instability problem incurred by inappropriate selection of the scale parameters. Furthermore, it needs no order statistics, such as the median, and hence no sorting. This improves the computational efficiency.

6.1.5 Convex Priors for DA and M-Estimation

Encoding the edge-preserving ability into prior distributions may lead to *nonconvex* energy functions. This is the case in many models such as the well-known line-process model (Geman and Geman 1984; Blake and Zisserman 1987). Models with nonconvex energy functions have two disadvantages. The first is the instability of the solution (i.e., the energy minimum) w.r.t. the data (Bouman and Sauer 1993; Stevenson et al. 1994). A small change in the input might result in a drastic difference in the solutions. The phenomenon is also due to a hard decision-making property of nonconvex models (Blake and Zisserman 1987). As such, the solution often depends substantially on the method used to perform the minimization. The second disadvantage is

the high computational cost associated with the solution-finding procedure. An annealing process, either deterministic or stochastic, is incorporated into a local search algorithm in order to locate the global minimum (Geman and Geman 1984; Blake and Zisserman 1987). This makes the minimization procedure inefficient.

There has been considerable interest in convex energy models with edge-preserving ability (Shulman and Herve 1989; Green 1990; Lange 1990; Bouman and Sauer 1993; Stevenson et al. 1994; Li et al. 1995). This class of models overcomes the problems mentioned above with nonconvex functions. First, in terms of defining the minimal solution, the convexity guarantees the stability w.r.t. the input and makes the solution less sensitive to changes in the parameters (Bouman and Sauer 1993). The second advantage is the computational efficiency in searching for the global solution. Because there is only one unique solution, gradient-based minimization techniques can be efficiently utilized. Time-consuming techniques for tackling the local minimum problem, such as continuation or annealing, are not necessary in convex minimization.

Shulman and Herve (1989) proposed to use

$$g(\eta) = \begin{cases} \eta^2 & |\eta| \leq \gamma \\ 2\gamma|\eta| - \gamma^2 & |\eta| > \gamma \end{cases} \tag{6.12}$$

for computing optical flows involving discontinuities. This is the error function used in Huber's robust M-estimator (Huber 1981). The function has also been applied to curve fitting (Stevenson and Delp 1990), surface reconstruction (Stevenson et al. 1994), and image expansion (Schultz and Stevenson 1994). It has been shown to be advantageous in terms of computational complexity and reconstruction quality (Stevenson et al. 1994).

Similar to the Huber function, the function of Green (1990)

$$g(\eta) = \ln(\cosh(\eta/\gamma)) \tag{6.13}$$

is approximately quadratic for small η and linear for large values. Lange (1990) suggested using

$$g(\eta) = \frac{1}{2}\left(|\eta/\gamma| + \frac{1}{1+|\eta/\gamma|} - 1\right) \tag{6.14}$$

which can be described by seven properties.

Bouman and Sauer (1993) construct a scale-invariant Gaussian MRF model by using

$$g(\eta) = |\eta|^p \tag{6.15}$$

where $1.0 \leq p \leq 2.0$. When $p = 2$, it becomes the standard quadratic function. When $p = 1$, the corresponding estimator is the sample median and allows discontinuities. The results show that edges are best preserved when $p = 1$ and deteriorated for $p > 1$. The reason will be given in the next section.

Stevenson, Schmitz, and Delp (1994) presentd a systematic study on both convex and nonconvex models and give the following four properties for a function g to have good behavior: (i) convex, (ii) symmetric (i.e., $g(\eta) = g(-\eta)$), (iii) $g(\eta) < \eta^2$ for $|\eta|$ large to allow discontinuities, and (iv) controlled continuously by a parameter γ. They define a class of convex potential functions

$$g(\eta) = \begin{cases} |\eta|^p & |\eta| \le \gamma \\ (|\eta| + (\frac{p}{q}\gamma^{p-1})^{\frac{1}{q-1}} - \gamma)^q & \\ \quad +\gamma^p - (\frac{p}{q}\gamma^{p-1})^{\frac{q}{q-1}} & |\eta| > \gamma \end{cases} \tag{6.16}$$

with three parameters: γ, p, and q. When $1.0 \le p \le 2.0$ and $\gamma = \infty$, it is the same as that used by Bouman and Sauer; when $p = 2.0$ and $q = 1.0$, it is the Huber error function. The values $p = 1.8$ and $q = 1.2$ are suggested in that paper. We point out that (iii) is too loose and inadequate for preserving discontinuities. An example is $g(\eta) = \eta^2 - g_0$, where $g_0 > 0$ is a constant; it satisfies (iii) but is unable to preserve discontinuities. As pointed out in the next section, it is the derivative of $g(\eta)$ that determines how a model responds to discontinuities.

6.1.6 Half-Quadratic Minimization

The AM-estimator with convex priors can be explained by half-quadratic (HQ) minimization. HQ performs continuous optimization of a nonconvex function using the theory of convex conjugated functions (Rockafellar 1970). Since the introduction of HQ minimization to the field of computer vision (Geman and Reynolds 1992), HQ has now been used in M-estimation and mean-shift for solving image analysis problems.

In HQ minimization, auxiliary variables are introduced into the original energy function. The resulting energy function then becomes quadratic w.r.t. the original variable when the auxiliary variables are fixed, and convex w.r.t. the auxiliary variable given the original variable (thus the name "half-quadratic"). An alternative minimization procedure is applied to minimize the new energy function. The convergence of HQ optimization is justified in (Nikolova and NG 2005; Allain et al. 2006). While HQ itself is a local minimization algorithm, global minimization via HQ can be achieved by applying the annealing M-estimator concept (Li 1996b).

Auxiliary Variables

Given a set of m data samples $d = \{d_i \mid 1 \le i \le m\}$, where $d_i = f + \eta_i$, the problem is to estimate the location parameter f under noise η_i. The distribution of η_i is not assumed to be known exactly. The only underlying assumption is that $\eta_1, \ldots, \eta_m$ obey a symmetric, independent, identical distribution (symmetric i.i.d.). A robust estimator has to deal with departures from this assumption.

Let the residual errors be $\eta_i = d_i - f$ and the error penalty function be $g(\eta_i)$, satisfying conditions of (5.27) and $g'(\eta) = 2\eta h(\eta)$. The M-estimate f^* is defined as the minimum of a global error function

$$f^* = \arg\min_f E(f) \tag{6.17}$$

where

$$E(f) = \sum_i g(\eta_i) \tag{6.18}$$

HQ minimization applies the theory of convex conjugated functions (Rockafellar 1970). For each term $g(\eta_i)$, introduce an auxiliary variable b_i and consider the dual function $G(b_i)$ of $g(\eta_i)$

$$G(b_i) = \sup_{\eta_i \in \mathbb{R}} \left\{ -\frac{1}{2} b_i \eta_i^2 + g(\eta_i) \right\} \tag{6.19}$$

$G(b_i)$ is convex w.r.t. b_i. We have reciprocally

$$g(\eta_i) = \inf_{b_i \in \mathbb{R}} \left\{ \frac{1}{2} b_i \eta_i^2 + G(b_i) \right\} \tag{6.20}$$

The infimum is reached at the explicit form (Charbonnier et al. 1997)

$$b_i = 2h(\eta_i) = \begin{cases} g''(0^+) & \text{if} \quad \eta_i = 0 \\ \frac{g'(\eta_i)}{\eta_i} & \text{if} \quad \eta_i \neq 0 \end{cases} \tag{6.21}$$

With the auxiliary variables $b = \{b_i\}$ in (6.18), we get

$$\tilde{E}(f, b) = \sum_i \left\{ \frac{1}{2} b_i (d_i - f)^2 + h(b_i) \right\} \tag{6.22}$$

The infimum of $\tilde{E}(f, b)$ with a fixed f is

$$E(f) = \min_b \{\tilde{E}(f, b)\} \tag{6.23}$$

The optimal configuration can then be represented as

$$f^* = \arg\min_f E(f) = \arg\min_{f,b} \tilde{E}(f, b) \tag{6.24}$$

Alternate Minimization

The new energy function (6.22) is quadratic w.r.t. f when b is fixed and convex w.r.t. b given f. So, it can be efficiently optimized by using an alternate minimization algorithm.

Given (f^{t-1}, b^{t-1}) as the solution at the $(t-1)$th step, the tth step calculates

$$b^t = \arg\min_{b\in R} \tilde{E}(f^{t-1}, b) \tag{6.25}$$

$$f^t = \arg\min_{f\in\mathbb{F}} \tilde{E}(f, b^t) \tag{6.26}$$

The solutions to these two minimization problems can be found analytically as

$$b_i^t = 2h(d_i - f^{t-1}) \tag{6.27}$$

$$f^t = \frac{\sum_i b_i^t\, d_i}{\sum_i b_i^t} \tag{6.28}$$

If we choose the $g(\eta)$ to be one of the robust functions given in Table 6.1, the alternate minimization reduces to

$$f^t = \frac{\sum_i h(d_i - f^{t-1})\, d_i}{\sum_i h(d_i - f^{t-1})} \tag{6.29}$$

which is (6.5). A convergence analysis of the alternate minimization can be found in (Nikolova and NG 2005).

The connection between the widely used mean-shift (MS) algorithm (Fukunaga 1990; Comaniciu and Meer 1997) and HQ optimization is explained in (Yuan and Li 2007). The MS algorithm maximizes the following kernel density estimation (KDE) w.r.t. the mean f:

$$p(d \mid f) = \sum_i w_i k((d_i - f)^2) \tag{6.30}$$

where k is the kernel function and the predetermined weights satisfy $\sum_i w_i = 1$. The MS algorithm is obtained immediately by solving the gradient equation of $p(d \mid f)$ via fixed-point iterations:

$$f^t = \frac{\sum_i w_i k'\left((d_i - f^{t-1})^2\right)\, d_i}{\sum_i w_i k'\left((d_i - f^{t-1})^2\right)} \tag{6.31}$$

By setting $g(t) = -k(t^2)$ in (6.18), we can see that maximizing (6.30) is equivalent to minimizing the error penalty function (6.18).

By applying HQ minimization to (6.30), we get the alternate maximization algorithm

$$b_i^t = -2k'((d_i - f^{t-1})^2) \tag{6.32}$$

$$f^t = \frac{\sum_i w_i b_i^t\, d_i}{\sum_i w_i b_i^t} \tag{6.33}$$

which is identical to the iteration in the MS algorithm and AM-estimator. This also explains the MS algorithm from the HQ optimization perspective.

Annealing HQ

The HQ algorithm can be used with an annealing schedule to approximate the global solution. Replacing the scale estimate in the HQ by a bandwidth parameter $\gamma > 0$ and using an adaptive potential function g_γ with the adaptive interaction function h_γ gives the energy function

$$E_\gamma(f) = \sum_i g_\gamma(d_i - f) \tag{6.34}$$

The alternate minimization is then

$$b_i^t = 2h_\gamma(d_i - f^{t-1}) \tag{6.35}$$

$$f^t = \frac{\sum_i b_i^t \, d_i}{\sum_i b_i^t} \tag{6.36}$$

The local minimum problem can be overcome using the idea of AM-estimation (Li 1996b). Denote f_γ^* as the local minimum obtained under γ. The annealing is performed by continuation in γ as

$$f^* = \lim_{\gamma \to 0+} f_\gamma^* \tag{6.37}$$

Initially, γ is set to a value large enough to guarantee that the corresponding $g_\gamma(d_i - f)$ is convex in an interval. With such γ, it is easy to find the unique minimum of the $E_\gamma(f)$ using HQ minimization. The minimum is then used as the initial value for the next phase of minimization under a lower γ to obtain the next minimum. This way, we track the global minimum for decreasing γ values and hopefully approximate the global minimum f^* under $\gamma \to 0^+$.

6.2 Experimental Comparison

Two experiments are presented. The first is a general comparison of two estimators, the AM-estimator and the M-estimator with Tukey's biweight function, with simulated data. The second deals with an application. Experimental results demonstrate that the AM-estimator is significantly better than the traditional M-estimator in estimation accuracy, stability, and breakdown point.

6.2.1 Location Estimation

Simulated data points in 2D locations are generated. The data set is a mixture of true data points and outliers. First, m true data points $\{(x_i, y_i) \mid i = 1, \ldots, m\}$ are randomly generated around $\bar{f} = (10, 10)$. The values of x_i and y_i obey an identical, independent Gaussian distribution with a fixed mean value of 10 and a variance value V. After that, a percentage λ of the m

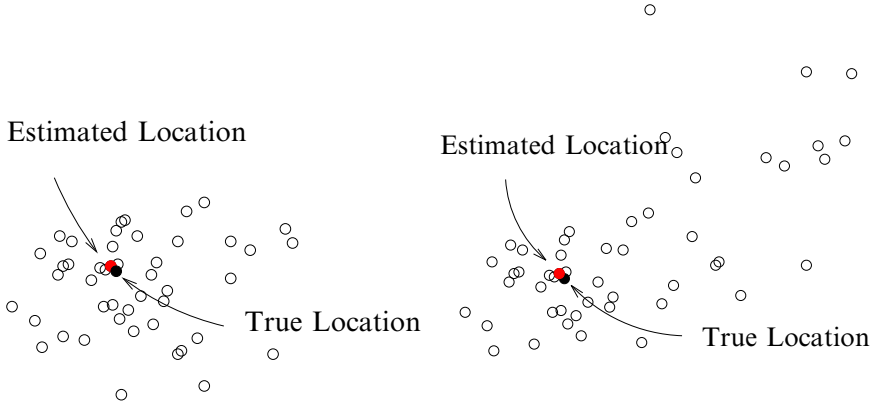

Figure 6.3: The AM-estimate of location. From (Li 1995a) with permission; ©1995 Elsevier.

data points are replaced by random outlier values. The outliers are uniformly distributed in a square of size 100×100 centered at $(b, b) \neq \bar{f}$. There are four parameters to control the data generation. Their values are:

1. the number of data points $m \in \{50, 200\}$,
2. the noise variance $V \in \{0, 2, 5, 8, 12, 17, 23, 30\}$,
3. the percentage of outliers λ from 0 to 70 with step 5, and
4. the outlier square centered parameter $b = 22.5$ or 50.

The experiments are done with different combinations of the parameter values. The AIF is chosen to be $h_{3\gamma}(\eta) = 1.0/(1 + \eta^2/\gamma)$. The schedule in lower$(T)$ is $\gamma \leftarrow \left(\frac{100}{t^2}\right)^{1.5} - 1$; when time $t \to \infty$, $\gamma \to 0^+$. It takes about 50 iterations for each of these data sets to converge.

Figure 6.3 shows two typical data distributions and estimated locations. Each of the two data sets contains 32 Gaussian-distributed true data points and 18 uniformly distributed outliers. The two sets differ only in the arrangement of outliers, while the true data points are common to both sets. The algorithm takes about 50 iterations for each of these data sets to converge. The estimated locations for the two data sets are marked in Fig. 6.3. The experiments show that the estimated locations are very stable regardless of the initial estimate, though the outlier arrangements are quite different in the two sets. Without the use of AM-estimation, the estimated location would have been much dependent on the initialization.

In a quantitative comparison, two quantities are used as the performance measures: (1) the mean error $\bar{e}$ versus the percentage of outliers (PO) λ and (2) the mean error $\bar{e}$ versus the noise variance (NV) V. Let the Euclidean

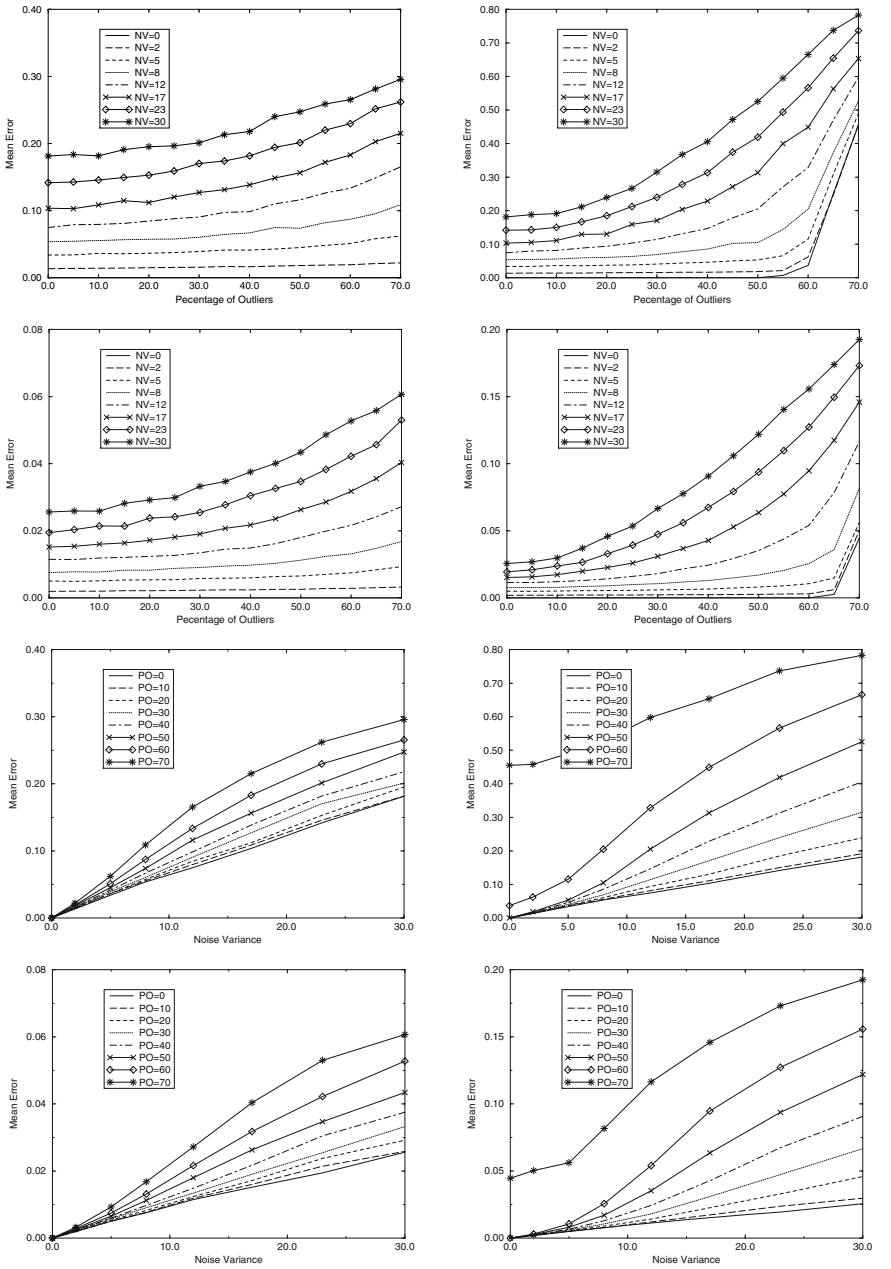

Figure 6.4: Mean error of the AM-estimate. From (Li 1996b) with permission; ©1996 Elsevier.

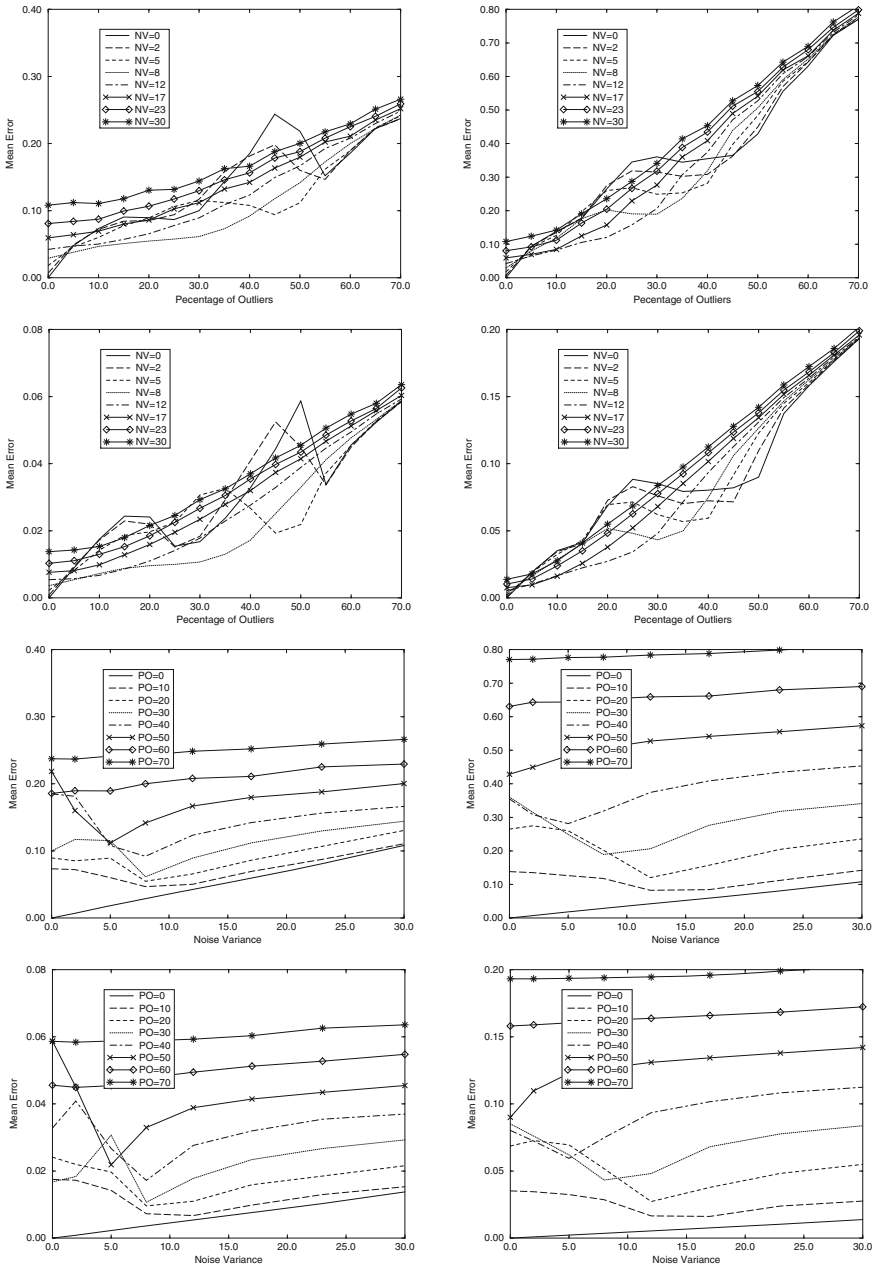

Figure 6.5: Mean error of the M-estimate. From (Li 1996b) with permission; ©1996 Elsevier.

error be $e = \|f^* - \bar{f}\| = \sqrt{(x^* - 10)^2 + (y^* - 10)^2}$, where f^* is the estimate and $\bar{f}$ is the true location.

Figures 6.4 and 6.5 show the mean errors of the AM-estimator and the M-estimator, respectively. Every statistic for the simulated experiment is made based on 1000 random tests, and the data sets are exactly the same for the two estimators compared. Outliers are uniformly distributed in a square centered at $b = 22.5$ (the left columns) or $b = 50$ (the right columns). The plots show the mean error versus percentage of outliers with $m = 50$ (row 1) and $m = 200$ (row 2) and the mean error vs. noise variance with $m = 50$ (row 3) and $m = 200$ (row 4). It can be seen that the AM-estimator has a very stable and elegant behavior as the percentage of outliers and the noise variance increase; in contrast, the M-estimator not only gives a higher error but also has an unstable behavior.

6.2.2 Rotation Angle Estimation

This experiment compares the AM-estimator with the M-estimator in computing the relative rotation of motion sequences. Consider the sequence of images in Fig. 6.6. Corners can be detected from these images as in Fig. 6.7

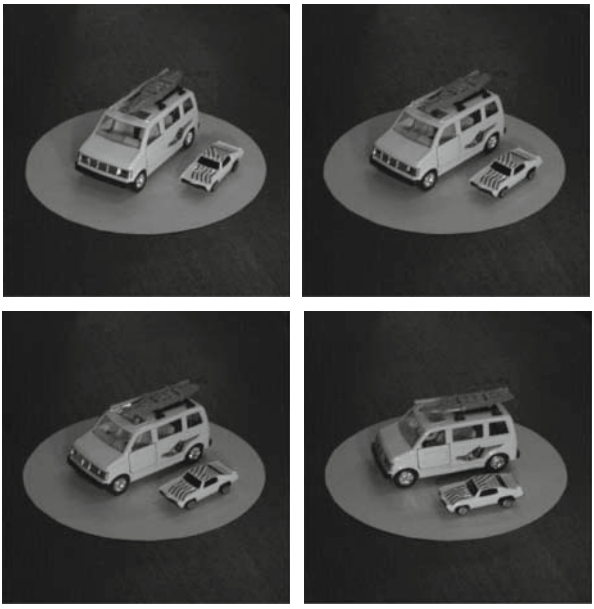

Figure 6.6: Part of a sequence of images rotating at 10 degrees between adjacent frames. The image size is 256×256.

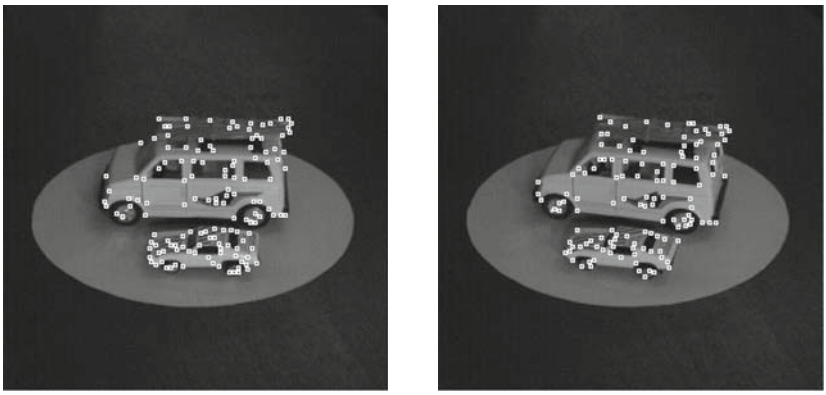

Figure 6.7: Corners detected.

by using the Wang-Brady detector (Wang and Brady 1991). The data

$$d = \{(p_i, p_i') \mid i = 1, \ldots, m\} \tag{6.38}$$

where $p_i = (x_i, y_i)$ and $p_i' = (x_i', y_i')$ represent a set of matched point pairs between two images. A previous work (Wang and Li 1994) showed that when the rotation axis $\ell = (\ell_x, \ell_y, \ell_z)^T$ is known, a unique solution can be computed using only one pair of corresponding points and the LS solution can be obtained using m pairs by minimizing

$$E(f) = \sum_{i=1}^{m} \{A_i \tan(f/2) + B_i\}^2 \tag{6.39}$$

where

$$\begin{aligned} A_i &= \ell_z(\ell_y(x_i + x_i') - \ell_x(y_i + y_i')) \\ B_i &= \ell_x(x_i - x_i') + \ell_y(y_i - y_i'). \end{aligned} \tag{6.40}$$

A unique solution exists for the LS problem (Wang and Li 1994). It is determined by the equation

$$\sum_{i=1}^{m} \left\{ [A_i \tan(f/2) + B_i] \cdot A_i/2 \cdot \sec^2(f/2) \right\} = 0 \tag{6.41}$$

where $f^* \neq 180°$; that is,

$$f^* = 2 \arctan\left(- \frac{\sum A_i B_i}{\sum A_i^2} \right) \tag{6.42}$$

The formulation above is based on an assumption that all the pairs $\{(x_i, y_i), (x_i', y_i')\}$ are correct correspondences. This may not be true in practice. For example, due to acceleration and deceleration, turning and occlusion,

the measurements can change drastically and false matches (i.e., outliers) can occur. The LS estimate can get arbitrarily wrong when outliers are present in the data d. When outliers are present, the M-estimator can produce a more reliable estimate than the LS estimator. The AM-estimator further improves the M-estimator to a significant extent.

The AM-estimator minimizes, instead of (6.39),

$$E(f) = \sum_{i=1}^{m} g_\gamma(A_i \tan(f/2) + B_i) \tag{6.43}$$

where g_γ is an adaptive potential function. By setting $\frac{\mathrm{d}E}{\mathrm{d}f} = 0$ and using $g'_\gamma(\eta) = 2\eta h_\gamma(\eta)$, one obtains

$$\sum_{i=1}^{m} \left\{ [A_i \tan(f/2) + B_i] \cdot h_\gamma(A_i \tan(f/2) + B_i) \cdot A_i/2 \cdot \sec^2(f/2) \right\} = 0 \tag{6.44}$$

Rearranging this equation gives the fixed-point equation

$$f = 2 \arctan\left(-\frac{\sum_{i=1}^{m} h_i A_i B_i}{\sum_{i=1}^{m} h_i A_i^2} \right) \tag{6.45}$$

where $h_i = h_\gamma(A_i \tan(f/2) + B_i)$. It is solved iteratively with decreasing γ values.

Figures 6.8 and 6.9 show the estimated rotation angles (in the vertical direction) between consecutive frames (the label on the horizontal axis is the frame number) and the corresponding standard deviations (in vertical bars) computed using the LS-, M-, and AM-estimators. From Fig. 6.8, we see that with 20% outliers, the M-estimator still works quite well while the LS-estimator has broken down. In fact, the breakdown point of the LS-estimator is less than 5%.

From Fig. 6.9, we see that the M- and AM-estimators are comparable when the data contains less than 20% outliers. Above this percentage, the

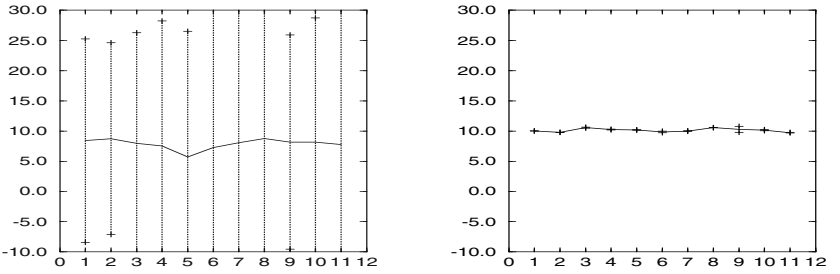

Figure 6.8: Rotation angles computed from the correspondence data containing 20% of outliers using the LS-estimator (left) and the M-estimator (right).

AM-estimator demonstrates its enhanced stability. The AM-estimator continues to work well when the M-estimate is broken down by outliers. The AM-estimator has a breakdown point of 60%. This illustrates that the AM-estimator has a considerably higher actual breakpoint than the M-estimator.

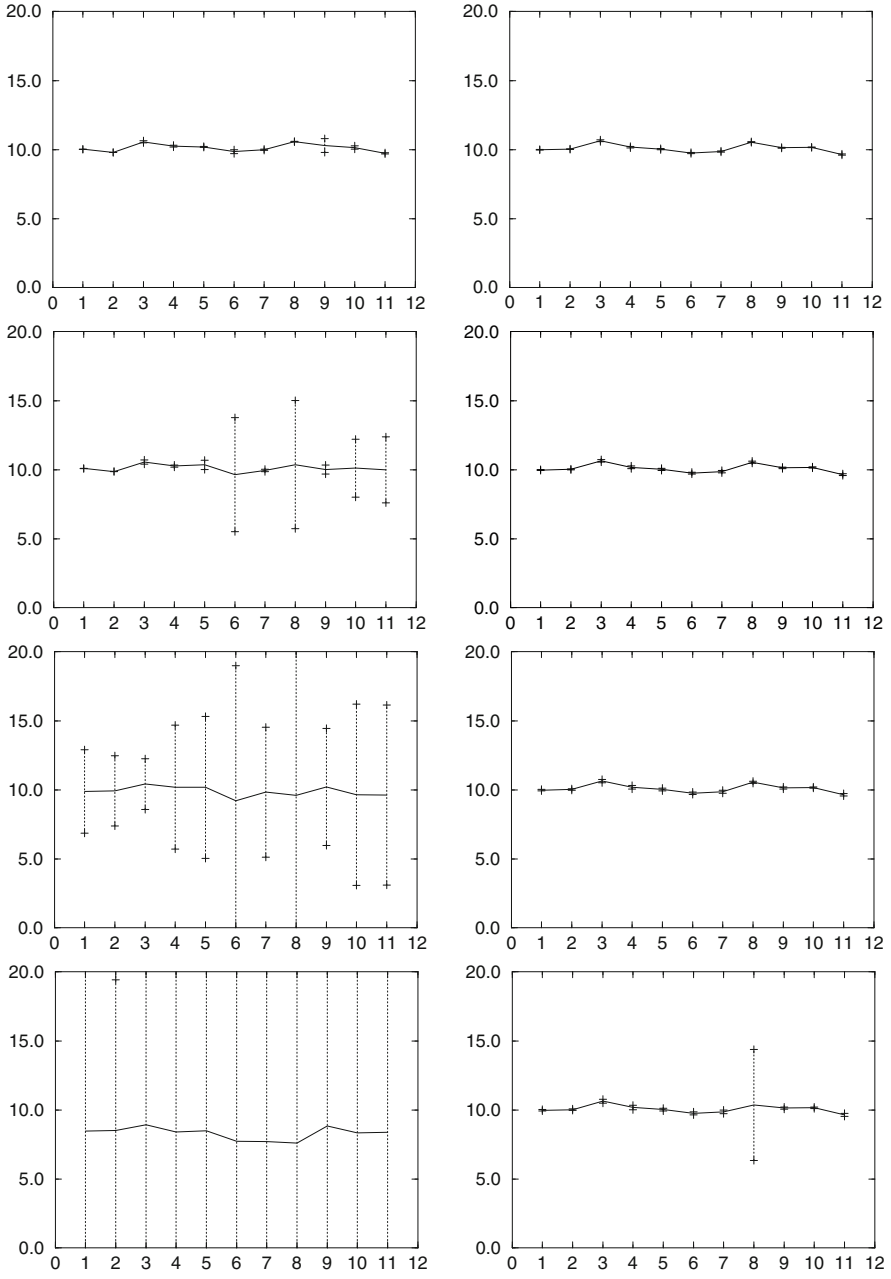

Figure 6.9: Results computed from the correspondence data containing 20% (row 1), 40% (row 2), 50% (row 3), and 60% (row 4) outliers using the M-estimator (left) and the AM-estimator (right).