

# Extracting Motion from Images: Robust Optic Flow and Structure From Motion

D. Suter, P. Chen and H. Wang

Dept. Elect. & Comp. Syst. Eng., P.O. Box 35, Monash University  
Clayton 3800, Vic. Australia: (d.suter,pei.chen,hanzi.wang)@eng.monash.edu.au

## Abstract

The first part of this paper summarises recent work in developing highly robust estimators and to apply these to computer vision applications - specifically for optic flow, range image fitting and segmentation, and image motion estimation.

The second part of the paper considers the so called “direct” methods for structure from motion (through matrix factorization), though the methods and principles are also applicable to other areas (face recognition, customer characterization and recommender systems etc.).

## 1 Highly Robust Fitting

Many of the tasks of computer vision can be cast as forms of statistical estimation and fitting. Because of the presence of multiple structures in the image, we need approaches that are robust to (pseudo)-outliers in the sense of having a high breakdown point. The two most used high breakdown point estimators are the Least Median of Squares [1, 2] (proved to be tolerant to up to 50% outliers but one can have configurations where it breaks down at a lower percentage) and RANSAC [3] (no theoretical breakdown point known). More recent techniques include ALKS [4], RESC [5] and MUSE [6]. Results presented in those papers *suggest* that these can resist substantially more than 50% outliers. For recent surveys, see [7] and [8].

These techniques usually employ random sampling techniques that aim to explore the search space of possible solutions well enough to have at least one candidate which is determined solely by inliers (to a single structure in the data). Secondly they have some form of model/fit scoring: median of the sorted residuals in the Least Median of Squares, and the number of residuals inside a certain chosen bound in RANSAC, for example. Wang and Suter [9, 10, 11, 12, 13]. have sought alternative ways of scoring candidate models, so that greater robustness may be achieved. [13, 10] employed symmetry in the data set: though somewhat limited in versatility, such an approach definitely restores robustness in situations

where standard Least Median of Squares will break down.

The purpose of this section is to describe more generally effective strategies that try to use more information from the residual distribution. In particular, we have used Kernel Density estimation and Mean Shift Techniques [14] to formulate model/fit scores that lead to empirically observed higher breakdown points than all existing methods.

The basic notion is that the robust estimate should produce a strong peak in the pdf of the residuals for that fit, and that the value of the residual corresponding to that peak should be small (ideally zero, of course). Maximization of the following objective function performs well:

$$MDPE = \frac{\left( \sum_{X_i \in W_c} \hat{f}(X_i) \right)^\alpha}{\exp(|X_c|)} \quad (1)$$

where the mean shift procedure [14] (see section 1.1) is used to find the mode of the residual density  $\hat{f}$  and to limit mean estimation to use only “inliers”  $X_i$  (with center value  $X_c$ ) within the mean-shift window  $W_c$ . This method is similar to the Residual Consensus (RESC) method [5]. Essentially that method estimates the pdf by using a histogram (whose bin size is chosen by compressing a large histogram of residuals using a heuristic procedure so that 12% of the residuals are found in the first bin). The RESC criterion is then:

$$RESC = \sum_i^m \frac{h_i^\alpha}{|r_i|^\beta} \quad (2)$$

where  $h_i$  is the histogram value and  $r_i$  the residual of the  $i$ 'th bin. The value  $m$  is chosen by another heuristic (4.4% of  $h_{max}$ ), so as to exclude outliers. (Note: in [5]  $\alpha = 1.3$  and  $\beta = 1$ .) In essence, the scores 1 and 2 differ in how one restricts attention to likely inliers (we use the mean shift window) and how one models the pdf (we use kernel density estimation *and* mean shift mode seeking). These differences lead to significant improvements in robustness.

In section 1.1 we describe more completely the algorithms we have developed. In section 1.2 we

show that approaches based upon these procedures can tolerate up to 90% or so of outliers (including pseudo-outliers) and outperform previous computer vision methods (MUPE, RESC, ALKS, RANSAC) and more widely known methods (Least Median of Squares and Least Trimmed Squares) in that regard.

## 1.1 Mean Shift Scoring

Although other kernels could be employed, we use the Epanechnikov kernel in its 1-D form:

$$K(x) = \begin{cases} \frac{3}{4}(1-x^2) & x^2 < 1 \\ 0 & otherwise \end{cases} \quad (3)$$

The kernel density estimator is then:

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x-X_i}{h}\right) \quad (4)$$

for a data set of  $n$  residuals  $X_i$  and using “bandwidth”  $h$ .

The mean shift is calculated by [14]:

$$\hat{f}(x) = M_h(x) = \frac{1}{n_x} \sum_{X_i \in S_h} X_i - x \quad (5)$$

where  $S_h$  is an interval of half-width  $h$  centred on  $x$  - the “mean shift window”.

Simple calculations show that iterating the above will converge to a local maximum of the estimated pdf. Because of the limited extent of the mean shift window (in the particular case of the kernel we choose, there is an inherent limit to the spatial influence of a data point, coming from the finite support of the Kernel - although a window could be imposed on other kernels not having finite support), this process is reasonably insensitive to outlier residuals.

Thus, the overall procedure is: employ random subset sampling (like Least Median of Squares etc.), rank the candidate subsets (via the residuals to the fit they determine) using the above mean shift procedure, iterated until the mode of the residuals (closest to 0) is located. Select the trial fit that maximises criterion (1).

## 1.2 Examples

### 1.2.1 Circle Finding

Although we acknowledge that one should consider using geometric distance, rather than residuals, in the minimization [15], here, we avoid the non-linear theory and resultant approximations of geometric fitting, and apply our procedure to the residuals produced by substituting the data points into the defining parametric form of the model.

We compare our method to the performance of the Hough transform, Least Median of Squares, RANSAC, ALKS, and RESC.

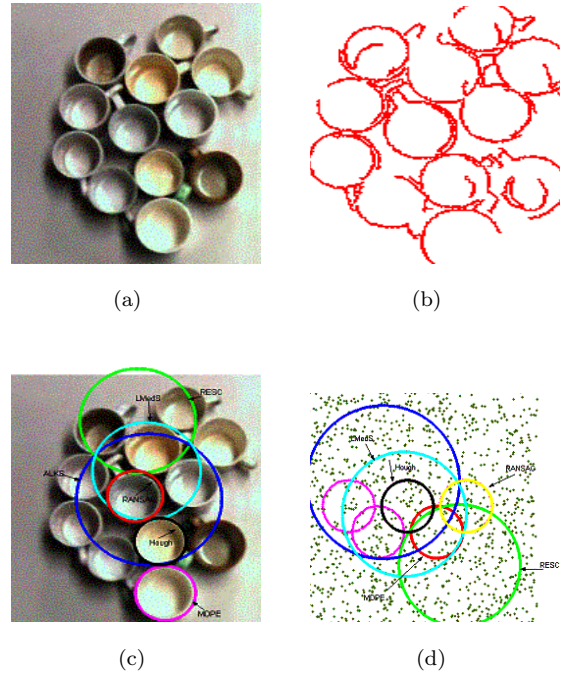


Figure 1: Circle finding

Examples of circle fitting can be found in Figure 1. The cups image 1(a) is processed with an edge detector 1(b). We obtain the results shown in 1(c). MDPE (pink), Hough (black) and RANSAC (red) all produce robust fits. Least Median of Squares (light blue), ALKS (dark blue) and RESC (green) all breakdown. As a second example, a synthetic data set was created to resemble “olympic rings” (pink - note, some rings are overwritten by the fitted rings - black, red and yellow) and then artificially corrupted with uniformly distributed noise samples - see Figure 1(d). In this example, the data that are “inliers” to any one ring constitute only about 5% of the data - that is, for a given fit, there are around 95% outliers to that fit. MDPE (red), RANSAC (yellow) and the Hough transform (black) can all find a ring. However, Least Median of Squares (light blue), ALKS (dark blue) and RESC (green) again breakdown.

These and other similar experiments we have performed, shows that algorithms based upon our MDPE criterion outperform other robust techniques (Least Median of Squares, techniques based upon least k’tth order statistics such as ALKS, and the RESC approach). The technique is challenged for robustness only by RANSAC (which requires a priori knowledge of the expected number of inliers) and the non-regression (limited precision) Hough transform (which also requires a well-chosen bin size for the parameter discretization).

The essential parameter required for our technique is the bandwidth of the kernel density estimator (al-

though there are the inevitable minor bounds and tolerance parameters that plague code in order to guard against certain numerical limits, or to decide when to cease iteration). In the experiments reported in this section, we empirically investigated the behaviour of our approach with varying bandwidth and found that overall performance, at least over a reasonable range of bandwidth choices, was reasonably stable. In the next section we present results where we have automatically chosen the bandwidth.

### 1.2.2 Optic Flow

Assuming that the imaged point at position  $(x, y)$  and time  $t$  maintains a constant brightness  $I(x, y, t)$ , as it moves under the optic flow  $(u, v)$ , then a simple differentiation reveals to first order:

$$\frac{\partial I}{\partial x}u + \frac{\partial I}{\partial y}v + \frac{\partial I}{\partial t} = 0 \quad (6)$$

- the ‘‘Optic Flow Constraint’’. Since this constraint provides only one constraint in two unknowns, it has been solved by assuming, within a small image patch, that the flow is of a simple parametric form (e.g., locally constant, or locally affine etc.).

Within each patch, we can measure the quantities  $\frac{\partial I}{\partial x}$ ,  $\frac{\partial I}{\partial y}$ , and  $\frac{\partial I}{\partial t}$ . However, since the measurements can contain very large errors, and since the patches may straddle different flows (moving according to different parameters), we need methods of solution that are robust to large numbers of outliers (including pseudo-outliers). Robust statistical methods such as Least Median of Squares have been demonstrated as superior to Least Squares and other competing methods in this respect [2].

We have applied our method, as described above but with some modifications, to the problem of optic flow estimation. Firstly, we have implemented a form of automatic selection of the kernel bandwidth. Using standard techniques [16]:

$$\hat{h} = \left[ \frac{243R(K)}{35u_2(K)^2n} \right]^{\frac{1}{5}} s \quad (7)$$

where  $R(K) = \int_{-1}^1 K(\xi)^2 d\xi$ ,  $u_2(K) = \int_{-1}^1 \xi^2 K(\xi) d\xi$  and  $s$  is the sample standard deviation. This still requires the estimation of the sample scale in a robust fashion and, since the above is the recommended upper bound on  $\hat{h}$ , we also have to employ a multiplicative adjustment  $c\hat{h}$  for  $0 < c < 1$ . Space does not permit us to go into details here, but we have investigated, with some success, both the use of order statistics for scale estimation and a rather more novel scheme of using a mean-shift like procedure to also find the pdf valley (thus providing a useful classification of inliers as those between the peak and the valley in the pdf). We have also, for speed reasons, modified the MDPE criterion to use what we

Technique	Ave Err
Bab-Hadiashar and Suter [2]	1.94
QMDPE ( $\sigma = 2.0, 25 \times 25, m=30$ )	1.34

Table 1: Yosemite Valley Sequence - Optic Flow

call QMDPE:

$$QMDPE = \frac{(\hat{f}(X_c))^\alpha}{exp(|X_c|)} \quad (8)$$

One sample of our results is provided in Table 1 where we compare with [2] (at this point the best competing robust method). Clearly QMDPE performs well.

## 2 Imputation and De-noising

There are many cases in computer vision where one *partially* observes a noisy version of a low rank matrix: structure from motion, face recognition, data modelling and visualisation etc. Thus we study the problem of recovering the ‘‘best’’ *complete* low rank matrix - imputing the missing values and denoising the matrix [17] [18].

More formally, a large matrix  $\mathbf{M}_{\text{true}} \in \mathbf{R}^{m,n}$ , which *should be of low rank*  $r$ , is observed ( $\mathbf{M}_{\text{obs}}$ ) through a process of i.i.d. Gaussian noise corruption of the entries and ‘‘occlusion’’ (term chosen partly because these may be track positions that become occluded).

### 2.1 Imputation

We seek the complete (‘‘not yet de-noised’’) matrix  $\hat{\mathbf{M}}$  that minimises the distance between itself and its rank- $r$  projection  $\|\hat{\mathbf{M}} - \hat{\mathbf{M}}^r\|_F^2$  subject to the recovered matrix having the same entries for the measured entries as those in the original measurement matrix. This differs from the similarly motivated objective of Shum/Wiberg [19] who seek  $\mathbf{M}_{\text{SW}}$  to minimise the measure:  $\|\mathbf{M}_{\text{obs}} - \mathbf{M}_{\text{SW}}\|_{F_{iw}}^2$  where  $F_{iw}$  is the obvious modification of the Frobenius norm to only include those terms where one has observations and to allow for weighting of the measurements in relation to some confidence measure.  $\mathbf{M}_{\text{SW}}$  is forced to be of rank  $r$  by expressing it as the product of an  $r$ -column matrix and a  $r$ -row matrix. We recover an imputed matrix that may then be optimally de-noised. The Shum-Wiberg combines the de-noising and imputation. Hartley and Schaffalitzky (*ibid*) discuss how to efficiently solve the Shum/Wiberg formulation. Such an approach also does not directly impute the ‘‘missing values’’ (they can be recovered through multiplication of the matrix factors).

We tackle our imputation iteratively. In a sense, our starting strategy has a similarity to the original Tomasi and Kanade [20] method in that we rearrange the data matrix so that it has a complete matrix in the upper left corner. We then address the question of how to add partially known rows and columns so as to “grow” the imputed matrix. *It is not necessary, though, that we find the arrangement that provides the largest such known submatrix to start with.* Nor, as stressed before, do we deliberately attempt to “de-noise” at this stage.

Starting from a complete submatrix (*not necessarily the largest one that we can find*), we grow by columns and rows by filling in the missing entries in the new row/column so that the new vector is as close as possible to the subspace spanned by the submatrix. This part is essentially the same as the first step of the SVD update of Brand [21].

However, the matrix when grown, is now closest to the original rank- $r$  approximation, but not necessarily closest to the rank- $r$  approximation of itself. Thus we iterate from here by obtaining the rank- $r$  approximation (SVD) and restart the matrix growing using this subspace. We can prove a convergence property for this iteration [17].

We call this algorithm “Iter”. In general, it outperforms Jacob’s method, produces approximately the same result as Shum’s method (when that method starts from a good initialisation) *when they both converge*. In the next section, by examining the gains in adding information, we describe a variant (“Iter-Part”) that is generally more successful and less prone to “wandering away from the solution” (though the latter happens relatively rarely with moderate amounts of missing data).

## 2.2 Denoising and Matrix Growth

It can be shown [18] that the denoising action of low rank projection has the following characteristic:

$E|A_{i,j} - A_{i,j}^r| = \sigma \sqrt{\frac{r(m+n)-r^2}{mn}}$  where  $\sigma$  is the scale of the noise. This quantifies the well known fact that as the size of the matrix grows relative to the rank, the resulting estimations should become more accurate. What is of more interest is whether (and to what extent) the above remains true when the growth comes about by adding rows and columns with missing values. There should be a trade off between increasing accuracy with more data and loss of accuracy with more missing data being imputed.

We note that the ratio in the above expression is the number of degrees of freedom in a rank- $r$   $m$  by  $n$  matrix, over the degrees of freedom in a general  $m$  by  $n$  matrix. We hypothesize that when there are  $p \gg r(m+n-r)$  observed values (and  $mn-p$  missing values), the error will behave as:

$E|A_{i,j} - A_{i,j}^r| = \sigma \sqrt{\frac{r(m+n)-r^2}{p}} \sqrt{\frac{1}{1-\rho}}$  for  $\rho = 1 - \frac{p}{nm}$ ,

the fraction of missing data. For reasonably small amounts of missing data, our experiments confirm that this is a reasonable approximation.

Thus we consider the following strategy: Try to use the most informative data as defined by the submatrix minimising  $\frac{r(m+n)-r^2}{p}$ . A heuristic method of achieving this is the best we can hope for as such problems are intractable. We simply take a “greedy” strategy of including rows (or cols) starting from those ordered with least missing values, until the ratio begins to increase.

## 2.3 Evaluation

Evaluation poses some problems. With synthetic data (complete ground truth known), one can evaluate by simply using the RMS error between the recovered (“reprojected”) values (missing and not) and the known values of the data matrix. With real image sequences, one doesn’t know the ground truth for the occluded points (although one can artificially occlude some points - pretend they are occluded). However, where there are genuinely occluded points, one generally disregards the reconstructed points in the error measure. Yet it is precisely these points that are the most challenging. One can gain some impression of the accuracy of these by plotting the “reprojected” points and looking for bizarre behaviour. When we have done so, for methods produced by other authors, we have found some rather unflattering results.

3D feature points were uniformly distributed in a cube  $[-500,500]*[-500,500]*[-500,500]$ . Different levels of Gaussian noise, with scale from 1 to 20, are added to the 2D feature points. Some algorithms, for any given run, may diverge. We detected divergence when the RMS error rises above 10 times the true error. We don’t include those divergent cases in the RMS index (Figure 2), of course. Thus to get a better idea of the performance, one also needs to look at the convergence rate for each algorithm (see Table 2).

We compare our two variants “Iter” and “Iter-Part” with three approaches that use Jacob’s method as the start (three variants: “rankrsf”, “rankrfsm-transpose” and “rank”) followed by Shum’s algorithm.

As an example of real data, we show (Figure 3) the results of reconstruction of the traces of the rotating dinosaur sequence.

## 3 Summary and Conclusions

### 3.1 Robust fitting

In this paper we have adopted the philosophy that a single statistic, such as the  $k$ ’th order statistic, or the number of inliers within a certain bound; is unlikely

	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
Iter	<b>100</b>	<b>100</b>	<b>99.90</b>	<b>99.95</b>	<b>99.90</b>	<b>99.45</b>	99.20	97.85	96.50	92.80
Shum1	<b>100</b>	<b>100</b>	<b>99.90</b>	<b>99.95</b>	99.80	99.30	98.75	96.15	91.20	81.50
Shum2	<b>100</b>	<b>100</b>	99.85	<b>99.95</b>	<b>99.90</b>	99.35	<b>99.40</b>	<b>97.90</b>	<b>96.70</b>	<b>93.45</b>
Shum3	99.90	99.95	99.80	99.65	99.55	99.00	98.60	95.80	92.45	85.55

Table 2: Convergence rates

Note: IterPart essentially converges all of the time - at least on these experiments.

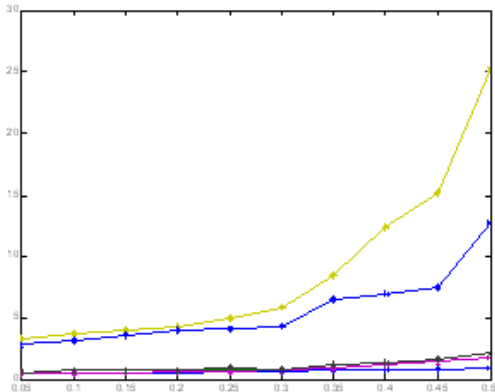


Figure 2: RMS Error - noise level 1

Graph of the average RMS reprojection error versus the percentage of missing data (up to 50%). From best to worst: “IterPart” is best, then four methods which are indistinguishable on this graph: “Iter”, “rankrsfm+Shum”, “rankr+Shum” and “rankrsfm-tpose”, then rankrsfm (slightly worse than the previous four), rankr and rankrsfm-tpose. The other noise levels created show similar trends.

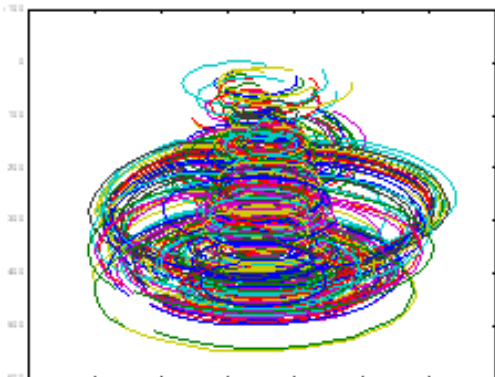


Figure 3: Rotating Dinosaur  
336 tracks recovered by our method - Jacob’s and Shum’s methods all produce wandering tracks.

to be a sophisticated enough measure to reliably discriminate between candidate model fits in a robust procedure that has to cope with a wide range of possible outlier/pseudo-outlier populations. Instead, we have looked at characterizing the quality of a model fit by more complex measures of the residual distribution: capturing information such as how peaked around zero the residual probability density function is. To this end, we have devised procedures that, at their core, use kernel density estimation of the pdf and a mean-shift approach to locate the peak of that pdf. Experiments have shown that such procedures can considerably out-perform existing robust techniques in terms of apparent breakdown point behaviour.

In concluding, we must remark on the shortcomings of the approaches we are hereby promoting. From a practical point of view, the methods are somewhat costly compared to Least Median of Squares and to RANSAC. Though we haven’t extensively studied the issue of just how computationally efficient the approaches can become: it would be optimistic to expect that they can compete with methods relying on much more simple measures of candidate solution quality. From a theoretical point of view, a lot remains to be studied. Though we promote our schemes in terms of “breakdown point”, we acknowledge a number of issues in respect of this. We have not formally defined “breakdown point”; nor, consequently, have we in any way attempted to prove attainment of a high breakdown point. In these respects, our approach is intuitive and empirical. However, we trust, despite these shortcomings, the techniques we have described will be of use to the computer vision community (and wider) as the basis of proven practical methods which can be refined, and whose theoretical underpinnings can be explored. Moreover, we must point out that, despite impressions that may be obtained by reading much of the literature, particularly that aimed more at the practitioner, more traditionally accepted techniques still have their shortcomings in similar ways. For example, though it is often cited that Least Median of Squares has a proven breakdown point of 50%, it is often overlooked that all practical implementations of Least Median of Squares are an approximate form of Least Median of Squares (and thus only have a weaker guarantee of robust-

ness). Indeed, the robustness of practical versions of Least Median of Squares hinges on the robustness of two components (and in two different ways): the robustness of the median residual as a measure of quality of fit and the robustness of the random sampling procedure to find at least one residual distribution whose median is not greatly affected by outliers. Our procedures, like many other procedures, share the second vulnerability as we too rely on random sampling techniques. The first vulnerability is sometimes disregarded for practical versions of Least Median of Squares, because robustness is viewed as being guaranteed by virtue of the proof of robustness for the ideal Least Median of Squares. However, two comments should be made in this respect. Firstly, that proof relies on assumptions regarding the outlier distribution and it can easily be shown that clustered outliers will invalidate that proof. Secondly, there is an inherent “gap” between a proof for an ideal procedure and what one can say about an approximation to that procedure. We believe that our method of scoring the fits better protects against the vulnerabilities that structure in the outliers expose. We have presented empirical evidence to support that.

### 3.2 Imputation and SFM

We have proposed an algorithm (with some variants) for recovering and factoring a noisy low rank matrix with missing values. We have shown that the algorithm works well, though we would stop short of claiming superiority. The algorithm is probably not competitive from a computational point of view (although the same can be said for some of the competitors). Perhaps of more value, is that we believe we have begun to understand the effects of imputation and the trade-off involved when one adds more data that carries extra missing values. We have proposed a scheme that seems to control that trade-off.

The algorithms and analysis we presented should be applicable to many other problems - we are currently looking at face recognition and related problems.

## References

- [1] P. Meer, D. Mintz, D.Y. Kim, and A. Rosenfeld. Robust regression methods in computer vision: A review. *International Journal of Computer Vision*, 6:59–70, 1991.
- [2] A. Bab-Hadiashar and D. Suter. Robust optic flow computation. *International Journal of Computer Vision*, 29(1):59–77, August 1998.
- [3] M. A. Fischler and R. C. Bolles. Random consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, June 1981.
- [4] K-M Lee, P. Meer, and R-H Park. Robust adaptive segmentation of range images. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 20(2):200–205, 1998.

- [5] X. Yu, T.D. Bui, and A. Krzyzak. Robust estimation for range image segmentation and reconstruction. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 16(5):530–538, 1994.
- [6] J.V. Miller and C.V. Stewart. MUSE: Robust surface fitting using unbiased scale estimates. In *Proc. Computer Vision and Pattern Recognition'96*, pages 300–306, June 1996.
- [7] P. Meer, C.V. Stewart, and D. Tyler. Robust computer vision: an interdisciplinary challenge. *Computer Vision and Image Understanding*, 78:1–7, April 2000.
- [8] C.V. Stewart. Robust parameter estimation in computer vision. *SIAM Reviews*, 41(3):513–537, September 1999.
- [9] H. Wang and D. Suter. A novel robust method for large numbers of gross errors. In *Proceedings ICARCV2002*, pages 326–331, 2002.
- [10] H. Wang and D. Suter. Using symmetry in robust model fitting. *Pattern Recognition Letters*, page to appear, 2003.
- [11] H. Wang and D. Suter. Variable bandwidth QMDPE and its application in robust optic flow estimation. In *Proceedings ICCV03, International Conference on Computer Vision, Nice, France*, to appear, 2003.
- [12] D. Suter and H. Wang. Robust fitting using mean shift: applications in computer vision. In *ICORS2003: International Conference on Robust Statistics, Antwerp, Belgium*, to appear, 2003.
- [13] H. Wang and D. Suter. LTSD: A highly efficient symmetry-based robust estimator. In *Proceedings ICARCV2002*, pages 332–337, 2002.
- [14] K. Fukunaga and L.D. Hostetler. The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Trans. Info. Theory*, IT-21:32–40, 1975.
- [15] K. Kanatani. *Statistical Optimization for Geometric Computation: Theory and Practice*. Elsevier Science, Amsterdam, The Netherlands, April 1996.
- [16] M.P. Wand and M. Jones. *Kernel Smoothing*. Chapman and Hall, 1995.
- [17] P. Chen and D. Suter. Recovering the missing components in a large noisy low rank matrix: Application to sfm. *submitted in revised form to IEEE PAMI*, 2003.
- [18] P. Chen and D. Suter. An analysis of linear subspace approaches for computer vision and pattern recognition. *submitted to SIAM Journal on Applied Mathematics*, 2003.
- [19] H-Y. Shum, K. Ikeuchi, and R. Reddy. Principal component analysis with missing data and its application to polyhedral object modelling. *IEEE PAMI*, 17(9):854–867, Sept. 1995.
- [20] C. Tomasi and T. Kanade. Shape and motion from image streams under orthography: A factorization method. *IJCV*, 9(2):137–154, 1992.
- [21] M. Brand. Incremental singular value decomposition of uncertain data with missing values. In *Proc. European Conference on Computer Vision ECCV'02*, pages 707–720, June 2002.

## Acknowledgment

Part of this work was carried out with the support of the Australian Research Council: grant A10017082. In particular, Wang was supported by that grant.