Chapter 1

FROM A ROBUST HIERARCHY TO A HIERARCHY OF ROBUSTNESS

Peter Meer
Electrical and Computer Engineering Department
Rutgers University
94 Brett Road
Piscataway, NJ 08854-8058, USA
meer@caip.rutgers.edu

How odd it is that anyone should not see that all observation must be for or against some view if it is to be of any service.

—Charles Darwin (1809-1882)

Abstract An attempt is made to present a (somewhat personal) history of how at the Computer Vision Laboratory in the late 1980's adopting robust techniques from statistics naturally arose from a quest for better multiresolution image analysis algorithms. While today these robust techniques are routinely used in the vision community, their rapid dissemination was in no small measure due to the unfettered research atmosphere which characterized the lab. Beside trying to record an instance of interdisciplinary research, a few technical issues (most of which were yet to be understood then) are also discussed.

Keywords: multiresolution image analysis; image pyramids; robust estimators; least median of squares; RANSAC; minimum volume ellipsoid

1. INSIDE IMAGE PYRAMIDS

When I arrived as a postdoc to the Center for Automation Research (CfAR) at the beginning of April 1986 and joined the pyramid group of the Computer Vision Laboratory, multiresolution image analysis was one of the main interests of Azriel. His edited volume (Rosenfeld, 1984) contained the most complete collection of papers dealing with hierarchical techniques in image analysis, and was the de facto textbook for this subfield. The "pyramid" group was very active, there were weekly meetings and a constant stream of technical reports (to become published papers soon afterward) was generated.

The research in the group was focussed on exploiting the multiresolution representation provided by an image pyramid to reliably delineate the significant features in an image. Azriel's technical report (Rosenfeld, 1986b) set clearly the main directions of investigation for the group. A short quote from the report best summarizes what were our goals.

> It should be pointed out that the pyramid techniques described in this paper are quite different from the ways in which pyramids have been used by other investigators (...). Pyramids are often used to generate a set of bandpass-filtered, sampled versions of an image. Our use of pyramids is quite different; we employ them for model *fitting* rather than for *filtering*.

For our purposes it suffices to describe only the simplest pyramid structure. Assume a $N \times N$, $N = 2^n$, input image at the base of the pyramid. A cell, or *parent*, on the l-th level of the pyramid is connected to an array of 2×2 cells, its *children*, on the $(l-1)$-th level. Thus, the height of the pyramid is $n = log_2 N$. In the bottom-up stage of an image analysis task the value of a parent is established based on the values of its children. The nature of the reduced resolution representation depends on the operation associated with generating the parents. At higher levels of the pyramid (lower resolution representations of the input) the features of interest are reduced to a few pixels and thus can be discriminated by local operations. A top-down tree growing process then recursively refines the extracted low-resolution representations and delineates the features of interest at the base of the pyramid, i.e., in the input image. The strong connection of the hierarchical image analysis paradigm to perceptual grouping was well documented (Rosenfeld, 1986a).

The above sketched hierarchical processing paradigm suggest two advantages relative to the traditional, single resolution techniques. A carefully chosen children→parent operation can yield very reliable low-resolution representations for the features of interest, representations which enhance these features against the background clutter. In order to have an effective noise reduction process, however, a necessary condition has to be satisfied. The representations of the features of interest must remain disjoint at *all* the levels of the image pyramid. Will return to this issue soon again since it is the starting point of our exposition. For tasks where the separability condition can be satisfied, the im-

age pyramid remains a valuable tool, as the achievements of the Sarnoff vision group show. See —— pointer here to Peter Burt's paper in the volume!!!!

The second advantage of image pyramids turned out to be less important. The hierarchical nature of the processing transforms an image analysis task which would require $O(N^2)$ operations on a single processor into an $O(logN)$ process on a cellular pyramid computer. However, such machines never became widely available, and by the early 90's the trend in computing turned against parallel computers on which pyramids could be emulated, e.g., (Sher and Rosenfeld, 1990). Today, in an era of 1Ghz personal computers, processing times are probably of lesser concern.

An important class of hierarchical algorithms we tried to develop was related to analyzing long, arbitrarily shaped features in an image. The difficulty of handling such features (ribbons, curves, region boundaries, etc.) was closely related to the limitations of image pyramids. Indeed, since these features could have any complex shape, at higher levels of the pyramid a cell could "see" several distinct fragments of it, without knowing that they belong to the same object at the input. It is easy to find an example where a curve is kept as single pixel fragments till the apex of the pyramid (Meer et al., 1990c), in which case the information reduction quality of the hierarchical processing is completely lost.

Earlier related work in the laboratory focussed on extraction of linear features, such as roads from aerial images (Shneier, 1982), or on hierarchical simulation of Gestalt laws, such as good continuation (Hong et al., 1983). In these applications the proposed algorithms only compensated for the rigid sampling structure of the image pyramid and were not general enough for our goal. Adaptive hierarchical structures, in which the resolution reduction is controlled by the local image content through reassignment of some child-to-parent links, were also tried for image segmentation, e.g., (Burt et al., 1981), (Peleg et al., 1986) and were subsequently investigated by other research groups as well (Spann et al., 1989), (Baronti et al., 1990).

In November 1988 Michel Bister arrived to CfAR as a visiting graduate student from the Vrije Universiteit, Brussels, Belgium. He was sent to learn more about image segmentation with pyramids and use the best available algorithm to analyze medical images. These images had ample fine structure and it was required that the segmentation preserves, as much as possible, the perceived topology of the input. In spite of best intentions (and a lot of work) Bister could not achieve a satisfactory segmentation with any of the algorithms available in the laboratory. He was then asked to perform an experiment in which the input image was shifted one or two pixels on the lattice before being segmented. That is, the relation between the image and the first pyramid level was slightly mod-

ified. While these small local changes were not expected to have a strong influence on the final segmentation, significant differences were observed among the obtained results. Bister concluded that the artifacts were intrinsic to any hierarchical image segmentation technique, and wrote up the results of the analysis in a paper (Bister et al., 1990) which received the best paper award for a publication in *Pattern Recognition Letters* in 1990.

By 1990 multiresolution image analysis already meant for most computer vision practitioners wavelets, scale-space, diffusion, etc. The book (Jolion and Rosenfeld, 1994) providing an excellent overview of what was achieved with image pyramids, thus in a way was also the epitaph of an era.

2. STOCHASTIC PYRAMIDS AND LEAST MEDIAN OF SQUARES

The search for accurate hierarchical representation of arbitrarily shaped objects lead to the introduction of stochastic pyramids (Meer, 1989). A stochastic pyramid is a hierarchy of graphs obtained by recursive graph contractions. The set of vertices retained for the next level of the hierarchy are chosen by local operations on the current level's graph, with a probabilistic algorithm breaking the tie situations. The local operations use the information extracted from the input and carried by the children of the vertex. Since the hierarchy is not restricted to a rigid sampling structure, it can mold itself to the features of interest. The approach was successfully applied to multiscale smoothing of chain-coded curves (Meer et al., 1990c) and segmentation of gray level images (Montanvert et al., 1991). Stochastic pyramids were also simulated on parallel computers (Ziavras and Meer, 1994), and the proposed hierarchical graph structure generated interesting theoretical questions e.g., (Kropatsch, 1995), as well as practical issues e.g., (Mathieu and Magnin, 1996).

Applying the stochastic pyramids to gray level image segmentation was triggered by a challenge posed by Walter Kropatsch during one of his short visits at CfAR in 1989. Walter questioned the ability of a hierarchical structure to efficiently manipulate information related to the topology of a complex binary image. The challenge was taken up by Annick Montanvert who was a visiting scientist with the pyramid group from the Joseph Fourier University in Grenoble. Annick had considerable expertise in digital geometry and soon we had implemented a hierarchical connected component delineation algorithm using the stochastic pyramid. Capturing the topology then became as simple as connecting the apexes of the individual graph hierarchies associated with the dif-

ferent connected components. Encouraged by this success we moved on to try to segment gray level images with a similar approach.

In the case of gray level image segmentation, the stochastic pyramid must generate a hierarchy of region adjacency graphs (RAG). To describe the resolution reduction process, lets assume for the moment that the input is homogeneous, i.e., all the pixels belong to the same region. The RAG of the input image is defined by the 8-connected graph of the underlying lattice. To generate the next level of the hierarchy only a subset of the vertices is retained, called the *survivors*. An optimal graph contraction process should be based only on local operations on the current level's graph. This can be achieved if the survivors satisfy the following two conditions on the graph:
(1) No two survivor vertices are neighbors.
(2) Any nonsurvivor vertex has a survivor neighbor.
The two conditions are equivalent with the vertices retained for the RAG of the next level being a maximal independent set of the RAG at the current level.

The survivors are selected with a parallel, probabilistic symmetric breaking algorithm (Meer, 1989). Every vertex in the graph is allocated a random number drawn from the $[0, 1]$ uniform distribution. A vertex becomes a survivor if its outcome is a local maximum and its neighbors on the graph are declared nonsurvivors for the subsequent iterations. After less than five iterations the algorithm converges and a maximal independent set of the graph is extracted. Note that the two necessary conditions on the survivors are automatically satisfied. The adjacency relations for the reduced resolution representation of the next level, i.e., the edges of that level's RAG, are obtained by using the paths between the survivors of the current level's RAG. Repeated runs of this probabilistic resolution reduction procedure, yield slightly different RAG hierarchies, however, a homogeneous region is always reduced in a logarithmic number of steps to a single vertex, the apex of the hierarchy.

To segment real images, first the current level's RAG must be decomposed into *similarity subgraphs* which correspond to homogeneous patches at the input. Given the parallel nature of the processing this has to be done exclusively by local operations. The above described resolution reduction procedure is then applied to each similarity subgraph and the results are combined into the next level's RAG.

The similarity subgraphs are defined using the *maximum averaged contrast* computed for each vertex. Let $g(v)$ be the gray level value associated with vertex v of an RAG and $g(v_i)$, $i = 1, \ldots, n_v$, the values associated with its neighbors on the graph. The attribute $g(v)$ was computed as a weighted average of its children's values. A local threshold $t(v)$ was derived by first computing the

differences

$$\delta_i(v) = \mid g(v_i) - g(v) \mid \qquad i = 1, \ldots, n_v \tag{1.1}$$

and then determining the largest jump between the left and right averages in the ordered sequence $\delta_{[i]}(v)$ for $j = 1, \ldots, n_v - 1$

$$L_j(v) = \frac{\sum_{i=1}^{j} \delta_{[i]}(v)}{j} \qquad R_j(v) = \frac{\sum_{i=j+1}^{n_v} \delta_{[i]}(v)}{n_v - j} \tag{1.2}$$

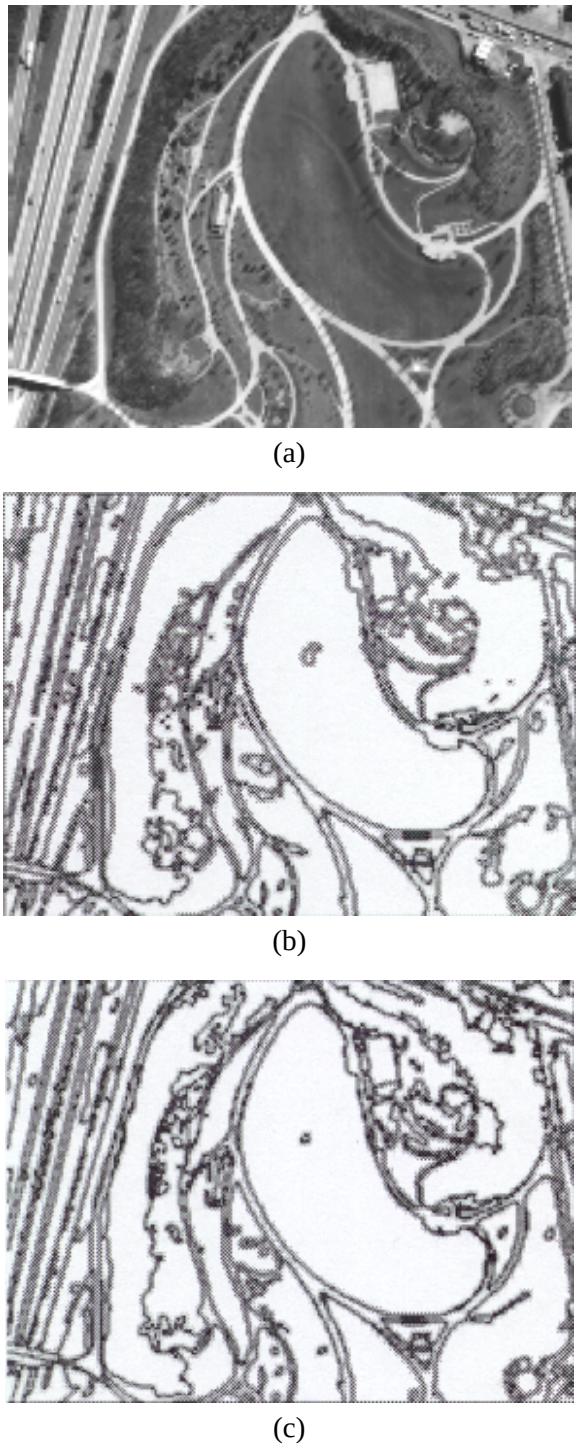
$$t(v) = \operatorname*{argmax}_j [R_j(v) - L_j(v)] \ . \tag{1.3}$$

The neighbors whose $\delta_i(v)$ was less or equal to the threshold were included into the same similarity graph as the vertex v. See (Montanvert et al., 1991) for a more detailed description.

The local operation employed to set the threshold was extremely simple and being based on averaging operations was certainly not robust. The obtained segmentations, however, were remarkable accurate and the probabilistic component did not seem to interfere with the delineation of the significant features. The image in Figure 1.1a has complex elongated regions whose boundaries are correctly recovered after the segmentation. The two different segmentations, in Figures 1.1b and c, obtained with two different RAG hierarchies differ only in the details where the implied piecewise constant image model is not accurate. The variability of the segmentation induced by the probabilistic component of the RAG hierarchy turned out to be more of an advantage than a drawback, and it was exploited to associate a confidence measure with the result. This research direction also has some connections with applying modern statistical techniques to computer vision tasks, and we will return to it in Section 3.

The quality of the segmentations was satisfactory since the RAG pyramid does not put any restrictions on the shape and the size of the delineated objects, as long as they have a contrast large enough to be associated with a distinct similarity subgraph. In the example in Figure 1.2 very fine features are delineated simultaneously with large homogeneous regions. Nevertheless, in our open ended quest for the "ultimate" hierarchical image segmenter it was just natural to ask ourselves if something would be gained when the local operation defining the similarity subgraphs becomes more sophisticated. This is how the issue of robust data analysis sneaked into the pyramid group of the Computer Vision Laboratory at CfAR.

The maximum averaged contrast is similar in spirit with Hinkley's cumulative sum (CUSUM) test for detecting jumps in the mean of a sequence of scalar

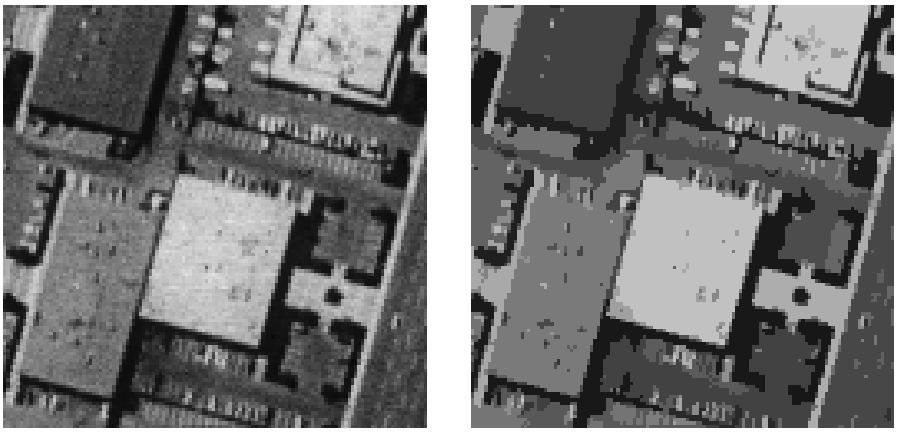


(a)

(b)

(c)

Figure 1.1 An example of image segmentation with RAG pyramids. (a) The input image. (b) and (c) The boundaries of the delineated regions for two different RAG hierarchies.



(a) (b)

Figure 1.2 Another example of image segmentation with RAG pyramids. (a) The input image. (b) Segmented image.

measurements (Basseville and Benveniste, 1983). It was clear from the beginning that a reliable local decision cannot be taken using only the values associated with the neighbors of the vertex and more data points will be required. We assumed that this can be assured by using also the sequences available to the children and if necessary the grandchildren of the vertex. However, once the question of discontinuity detection was separated from the process of building the RAG hierarchy, soon the effort was entirely focused on recovering the underlying structure of a piecewise linear, noisy one-dimensional signal.

The new research direction, while arose from pyramid related activities, was clearly moving away from the main themes of the pyramid group. Nevertheless, with Azriel's full support resources were allocated to it and the investigation started in earnest.

We began with the classical Chow test (Chow, 1960). This test can only be applied to one-dimensional sequences, and starts by fitting to the left and right of a data point a polynomial (linear) model using least squares. The residuals of these two fits, as well as of the fit to the combined region are used to build an F-type statistics. We implemented a multiscale approach with ad-hoc pruning of the discontinuity point candidates. While the results were satisfactory for synthetic data, the algorithm could not deliver reliable results for real data, such as a row of an image.

Dong Yoon Kim visited CfAR for a year in 1988 and 1989 on a fellowship from the Korean government. He had a doctorate in applied statistics and was very knowledgeable with recent developments in statistics where the issue of

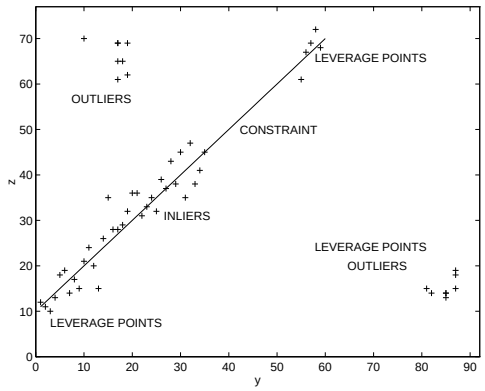


Figure 1.3 Definition of different measurements in relation to the assumed model.

robustness was the hot topic. The fact that Dong Yoon also had a background in computer science and could describe modern statistics in a simple, pragmatic way, certainly should be mentioned.

When the difficulty of decomposing the one-dimensional sequences into homogeneous segments came up in an informal discussion, Dong Yoon proposed to try robust estimation techniques. We were aware of the existence of M-estimators (see below) since they just started to to penetrate the image processing literature, e.g., (Kashyap and Eom, 1988), (Hansen and Chellappa, 1988). However, given the complexity of those applications, it was difficult to discriminate the role of M-estimators in the reported performance improvement.

Before proceeding with the history further, will present in a nutshell the two main robust techniques popular today in computer vision: M-estimators and the least median of squares (LMedS) type methods. Considering the simplest estimation problem suffices. Assume that $1-\epsilon$ percentage of the available measurements $z_j,\ j = 1, \ldots, n$ satisfy

$$z_j = \theta_1 + \theta_2 * y_j + \delta z_j \tag{1.4}$$

where δz_j is a zero-mean error term with unknown variance σ^2. Beside the first two moments and the independence of the measurement errors, no other assumption is required about the noise. The $\epsilon < 0.5$ percentage of remaining measurements do not obey (1.4), i.e., they are *outliers* relative to this model. Note that the outliers can still satisfy a relation such as (1.4) but with a different set of parameters (θ_1', θ_2').

In a rather simplistic definition, robust estimators are statistical techniques which can tolerate the presence of outliers in the data. The breakdown point of a robust estimator is the minimal percentage of contamination ϵ^* that renders

the estimate unreliable, i.e., its value becomes controlled by the outliers. The breakdown point, capturing the influence of the entire data set, characterizes the *global robustness* properties of the estimator.

Among the outliers we must distinguish an important subclass: the leverage points. In Figure 1.3 both the outliers and the leverage points are marked. By definition the leverage points are data points which have increased influence on the estimation process. They are not necessarily outliers, points at the boundaries of the data set are usually leverage points. However, points which are both outliers and leverage points (as those in the lower right corner) have an especially detrimental effect on the estimation process. For example, such points are not tolerated by the M-estimators.

An M-estimator is defined as the solution of the minimization problem

$$\hat{\boldsymbol{\theta}}_M = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \sum_{j=1}^{n} \rho(z_j; \boldsymbol{\theta}) \tag{1.5}$$

subject to the constraint assumed for the inliers, i.e., $z_{jo} = \mathbf{x}_{jo}^{\top}\boldsymbol{\theta}$, where the subscript o stands for uncorrupted (true) values, and the components of the *carrier* vector $\mathbf{x}_{jo}$ are known functions of the measurements. The objective function $\rho(u)$ must be positive valued and even symmetric with a unique minimum at the origin. It should be also nondecreasing for $u \geq 0$ and have piecewise continuous first two derivatives. Different choices of $\rho(u)$ yield different M-estimators. Note that $\rho(u) = \frac{1}{2}u^2$ is the traditional least squares estimator. The minimization problem (1.5) is most often solved iteratively with a weighted least squares procedure. The review paper (Li, 1985) is still the best practical reference for all the related topics.

It can be shown that all M-estimators have poor global robustness properties. Their breakdown point is zero since they cannot tolerate leverage points which are also outliers. The example in Figure 1.4 shows such a situation. Nevertheless, M-estimators are frequently and successfully used in computer vision. The reason is that extreme leveraging only rarely can appear in a vision task since the domain of definition of the variables is most often rather compact. Only the generalized M-estimators (GM-estimators), also known as bounded influence estimators have a nonzero breakdown point whose value is reciprocal to the dimension of the space of the unknown parameter $\boldsymbol{\theta}$.

The M-estimators have, however, an important advantage relative to the high breakdown point methods to be described below. They have excellent *local robustness* properties. That is, an infinitesimal change in the input data can lead to only an infinitesimal change in the output, i.e., the parameter estimate. In vi-

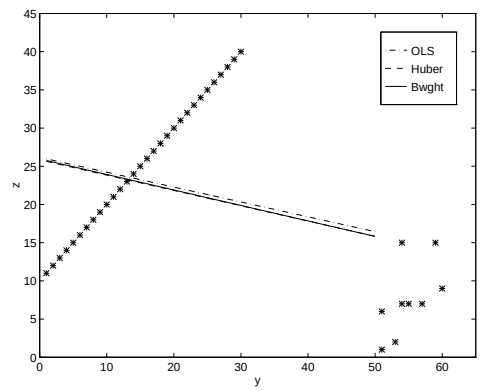


Figure 1.4 A simple example illustrating the failure of M-estimators in the presence of outlier/leverage points. OLS is the traditional least squares solution, Huber and biweight are two M-estimators which gave almost identical, and incorrect solution.

sion applications where the amount of outliers present in the data is small, such a stability may be more important than the capacity of rejecting all the outliers.

Many of the above mentioned properties of M-estimators were not well understood (at least outside the statistical community) in the late 80's. We did some experiments for decomposing the one-dimensional waveforms and the results were definitely better than using the Chow test. Nevertheless, the M-estimators based approach was rapidly abandoned when we became aware of a second family of robust estimation methods, that of the least median of squares (LMedS) type techniques.

The LMedS estimator solves

$$\hat{\boldsymbol{\theta}}_{LMedS} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \underset{j}{\operatorname{med}} \left(z_j - \mathbf{x}_{jo}^T \boldsymbol{\theta}\right)^2 \tag{1.6}$$

with the help of elemental subsets. An elemental subset is a p-tuple of randomly chosen data points from which a candidate of the parameter estimate can be uniquely determined. (The parameter vector $\boldsymbol{\theta}$ has dimension p.) The value of the median of the squared residuals is then computed and stored. Repeating the procedure several times, the LMedS estimate corresponds to the p-tuple which yielded the smallest value, i.e., the best approximated the objective function (1.6). The number of required trials can be determined from probabilistic considerations by allowing a very small chance of not finding a satisfactory solution. This number never exceeds a few thousand.

The LMedS estimator as presented above solves a regression problem. The minimum volume ellipsoid (MVE) estimator is its counterpart for multivariate location problems. Using again elemental subsets, the region of highest den-

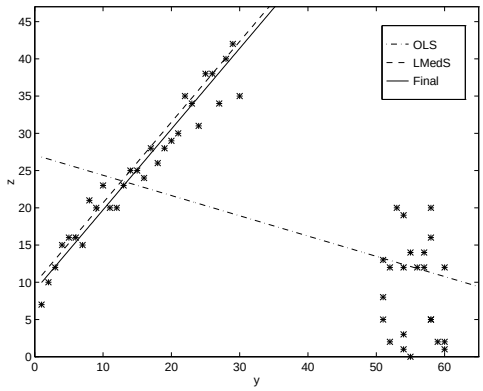


Figure 1.5 An example of the LMedS estimator handling outlier/leverage ponts.

sity in a feature space is located by identifying the smallest ellipsoid containing at least half the data points. This ellipsoid is then inflated to delineate the elliptical (normal) cluster associated with the inliers. The book (Rousseeuw and Leroy, 1987) has a complete treatment of all the topics related to least median of squares, and much more.

The LMedS family has excellent global robustness, its breakdown point being close to $\epsilon^* = 0.5$. In Figure 1.5 an example is shown in which the presence of a significant percentage of outlier/leverage points has no effect on the outcome of the estimation.

However, the LMedS family of estimators (and similar high breakdown point techniques) have poor local robustness properties. It is relative easy to find examples in which an infinitesimal change in the data drastically alters the output. In Figure 1.6 such an example is shown where a small change in a close to bimodal data can render the performance of the LMedS estimator similar to that of least squares.

The LMedS estimator also returns a scale estimate for the inlier noise, i.e., a quantity proportional with the estimated inlier noise variance. The scale estimate is the value of the minimization criterion (1.6) for $\hat{\boldsymbol{\theta}}_{LMedS}$. The inliers and outliers in the data then can be separated by examining the residuals in relation to a threshold derived from the scale estimate. Performing a least square postprocessing on *all* the inliers is recommended to improve the local robustness behavior of the LMedS estimator. These results are labeled 'Final' in Figures 1.5 and 1.6. Note that, as expected, the postprocessing cannot recover from the failure of the LMedS estimator.

A small confession is in order now. In the statistical literature the least median of squares estimator is known by the acronym LMS. So why did we change

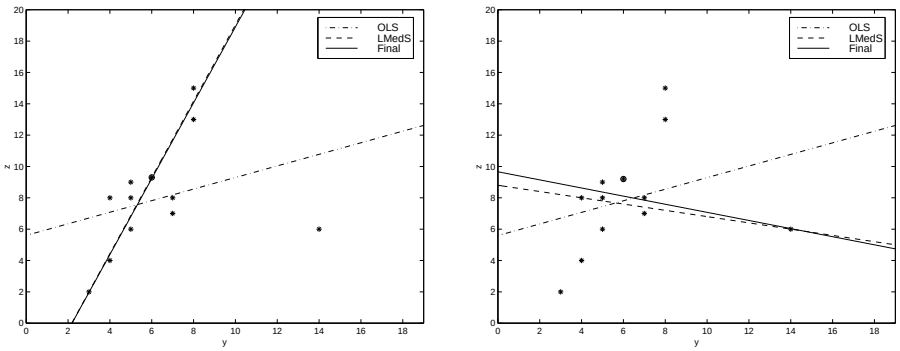


Figure 1.6 An example of the poor local robustness properties of the LMedS estimator. The only difference between the two data set is moving the point $(6, 9.3)$ in the left case to $(6, 9.2)$ in the right case.

its name to LMedS soon after our first conference publications? We were well aware that in the engineering estimation literature (including image processing) LMS stands for least mean squares, and did not want to create any confusion. Today in the vision literature both acronyms are used, often with interesting effects when results from (Rousseeuw and Leroy, 1987) are cited using LMedS as notation. To reveal the whole truth, for a moment we did consider defining the acronym as **L**east **Med**ian of **S**quares (LMdS), but found it a little too tacky for obvious reasons...

By the Fall of 1988 the work on applying the LMedS estimator to *any* low level vision task begun at earnest. Azriel decided that two students, Doron Mintz who was a Ph.D. candidate and John Kim who was an M.S. student, should dedicate most of their efforts to this topic. While we immediately realized the importance of this new tool for solving vision problems, it was yet not clear what can be achieved with it. We were also not aware that others were also discovering the potential of robust statistics.

Given its similarity to weighted least squares, it is not surprising that the M-estimators became popular before LMedS. Arguably the first two publications on using robust estimators in computer vision are the two conference papers in late 1988, (Besl et al., 1988) and (Haralick and Joo, 1988), and both employed M-estimators. In (Besl et al., 1988) a rather complex hierarchical scheme was proposed to adaptively model the local image structure by a low-order polynomial surface whose order was determined by analyzing the residuals of M-estimators. In (Haralick and Joo, 1988) the M-estimators were employed to improve the performance of pose estimation in the presence of erroneous matches. The vision group of the Artificial Intelligence Laboratory at the University of

Michigan, Ann Arbor, was the only one also experimenting with the LMedS estimator (Tirumalai and Schunk, 1988).

The first opportunity to publish our preliminary results came soon since the deadline for the proceedings of the 1989 DARPA Image Understanding Workshop was in early Spring. It was decided that a large part of the pages allocated to CfAR should be used to present the robust paradigm to the vision community. The paper came out to be 18 single spaced pages and had two goals: to serve as a tutorial for nonstatisticians on robust techniques, and to show what have we achieved using the LMedS family of estimators (Kim et al., 1989). All the experiments were based on simple synthetic data, but we have presented both regression and clustering results. Thanks in part to the high visibility of the workshop, and certainly due to the tutorial quality of the paper, the gospel of robustness began to spread rapidly in the vision community.

The 1989 IUW was held in May in Palo Alto, CA, and by that time it was undeniable that robust estimators will have a big impact on the field of image understanding. To facilitate the exchange of ideas Azriel and Larry Davis quicky decided to organize a workshop at CfAR on "Robust Estimation Techniques for Computer Vision". In retrospect, it seems that people were less overcommitted those days, since all the ten invited speakers were happy to participate on a short notice. The workshop took place on July 25 and 26, 1989 in College Park with 59 attendees. Beside the authors whose paper was already mentioned above, image processing related applications of the groups lead by Rama Chellappa (then at USC), Max Mintz and Saleem Kassam (both from UPenn) were also presented. Instead of a proceedings every speaker distributed paper at the meeting. This was certainly the first workshop in the vision community dedicated to robust methods, probably one of the first in which statistics was such a central topic. The next, larger international gathering on robust computer vision was held more than a year later in October 1990, in Seattle, organized by Bob Haralick and Wolfgang Förstner. By the late 1992 robust estimators became mainstream techniques in computer vision.

3. THE VISION PERSPECTIVE OF ROBUSTNESS

In one of the first discussions we had about the least median of squares, Azriel casually remarked that it reminded him of RANSAC. The RAndom SAmple Consensus was at that point a little known method, proposed in 1981 for pose estimation and model fitting (Bolles and Fischler, 1981), (Fischler and Bolles, 1981). Since except Azriel nobody heard about it, the lead was not followed up. We were, however, strongly called upon the fact again when the paper version of our research was submitted to the *International Journal on Computer*

Vision. One of the reviewers bitterly complained about not giving more attention to RANSAC beside mentioning it and in the final manuscript an entire section was dedicated to its relation to LMedS (Meer et al., 1991).

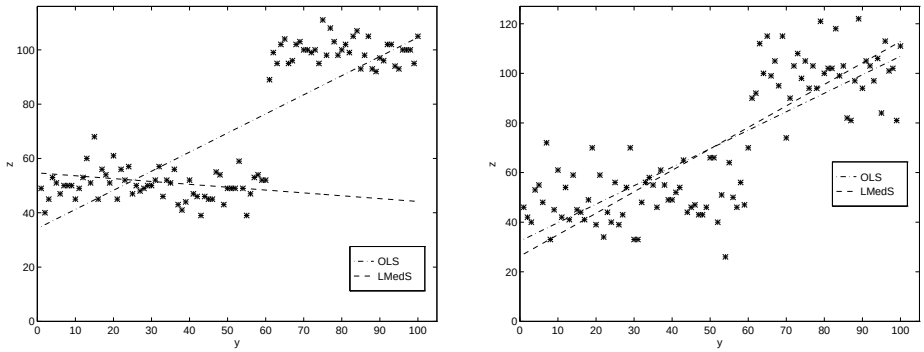
The similarity of RANSAC and LMedS is actually more on the computational than the theoretical level, but that was less clear a decade ago. Both use elemental subsets to obtain the estimate candidates. However, in the case of LMedS the median of the residuals squared is minimized, while in the case of RANSAC the number of points within a *given* tolerance from the fit is maximized. LMedS does not require an a priori scale estimate, RANSAC does. The advantage of a tuning parameter for RANSAC was recognized only relative recently. Since most often in computer vision applications the geometry of the problem allows to assess a reliable threshold for the inlier/outlier dichotomy, RANSAC can be adapted to the problem at hand. To deliver a reliable estimate LMedS by definition requires at least half of the data to be inliers, with RANSAC this condition can be relaxed.

Today, whenever a high breakdown point estimator is needed in solving a vision task, more and more often RANSAC (under many disguises) is employed instead of LMedS. It is somewhat paradoxical that one of the consequences of "importing" a valuable tool from statistics was the rediscovery of a technique proposed in our own field even before 1984 when LMedS appeared in statistics.

Another unexpected results was the poor performance of LMedS in the task for which was initially intended for: image segmentation. The segmentation employed the facet model, i.e., a piecewise polynomial surface representation of the gray levels. The input image was nonoverlappingly tessellated with windows in which LMedS was employed to find the zero or first order facet representing the inliers. This was followed by robust region growing across the boundaries of the tessellation to delineate the homogeneous regions. To say the least, the method was cumbersome, see (Meer et al., 1990b). It had numerous "patches" to make the region growing compensate for the known limitation of LMedS: the returned fit always represents what the estimator considers to be the absolute majority of the pixels in the window.

In spite of all the additions, the segmentation was clearly failing in some image regions where the LMedS window operator should have provided satisfactory performance. Furthermore, the quality of the segmentation was always inferior to the output of an RAG pyramid based delineation, the method discussed in Section 2. Recall that all the computational modules in the latter are nonrobust, simple local decisions. We did not expect this!

After some investigation the problem was isolated and reduced to its bare essence as shown in Figure 1.7. To have LMedS fail the data should be close to bimodal and significant noise present. Note that both the inliers and outliers



(a) (b)

Figure 1.7 An example of LMedS failure. (a) In moderate noise the inliers are correctly recovered. (b) In large noise there is no qualitative difference between the least squares and LMedS estimate.

obey the same model, only with different parameters, a situation typical to image structures. The inliers are the points centered around 50 and this is what the LMedS estimator should return. Nevertheless, once the measurement noise becomes large in relation to the inlier outlier separation, the LMedS estimator prefers a fit very similar to what the nonrobust least squares method yields (Figure 1.7b). Note that the two "clouds" of points are not overlapping, and when represented as a neighborhood in an image the two noisy surfaces were clearly distinguishable by eye, to our great frustration.

Extensive Monte Carlo simulations have shown that this erroneous behavior is intrinsic to the method and once the conditions are satisfied will always appear. The reasons for it can be understood examining Figure 1.8. Assume that the true fit is also available, i.e., was chosen by one of the elemental subsets. The distribution of the absolute values of the residuals are shown in Figure 1.8b for both the true fit and the LMedS fit which is the fit found as minimizing the optimization criterion (1.6). The former is bimodal the latter is unimodal with a long tail. Since the median selects the 50th percentile of these distributions, the presence of bimodality introduces a severe bias which makes the LMedS estimator to prefer the incorrect fit.

Some years later it was shown that similar problems appear with all the robust methods employed in vision applications (Stewart, 1997). Since data as in Figure 1.7 is not typical for statistics, the robust techniques developed there are not suitable to process it. Fortunately, such bimodal data with structured outliers only rarely appear in computer vision beyond low level tasks. One may conclude that for low level vision successful robust approaches not necessarily should take verbatim estimators from statistics. The Hough transform which

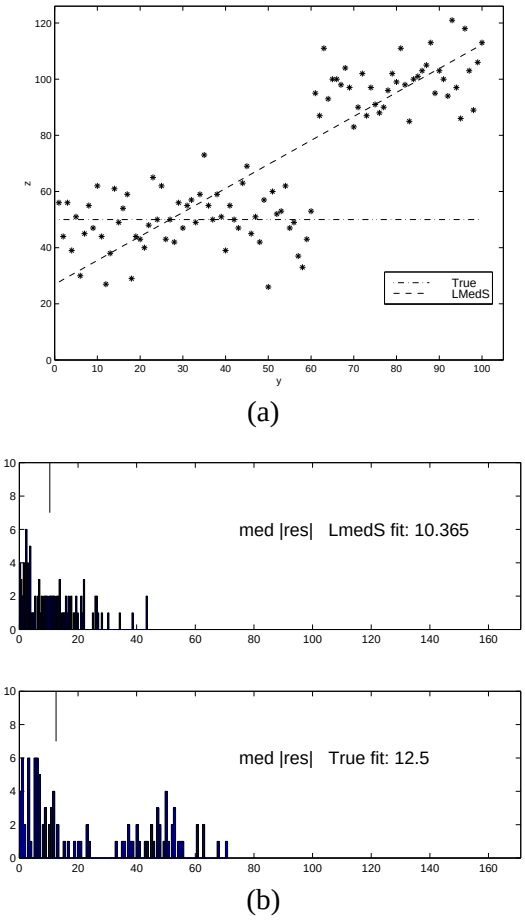


Figure 1.8 The mechanism of LMedS failure. (a) Critical data with the LMedS fit and the true values of the inliers shown. (b) The distribution for the absolute values of the residuals for the LMedS fit (top) and the true fit (bottom).

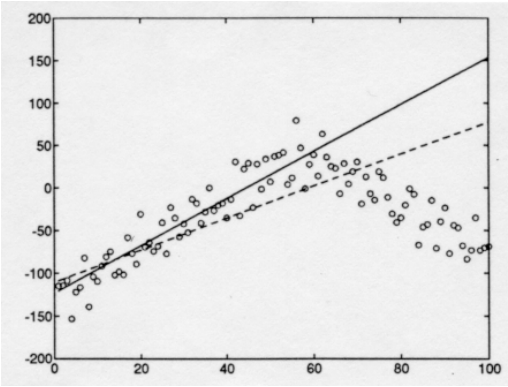


Figure 1.9 An example of applying the CBD procedure. The close bimodal piecewise linear data is analyzed with a linear model. The LMedS estimator (dashed line) fails to correctly identify all the inliers, the CBD procedure (solid line) does.

can also be interpreted as a multiple M-estimation, e.g., (Kiryati and Bruckstein, 1992), is probably the best example for an indigenous robust technique with superior qualities.

Once the failure of LMedS was understood we succeeded to design a procedure to avoid it. The procedure was called *consensus by decomposition* (CBD) and integrated two ideas. First, was the observation that the highest order coefficients of a polynomial surface are invariant under the translation of the coordinate system. Second, if the measurement noise is isotropic this property can be exploited to improve the performance of the zero-order LMedS estimator (mode seeking) for data such as in Figure 1.7. The CBD procedure in a window recursively estimated the coefficients of the polynomial surface starting from the highest order, reducing the estimation at each step to a mode seeking step.

The CBD approach was an empirical attempt to improve the performance of LMedS in the presence of difficult data. It did work satisfactorily, see the example in Figure 1.9, and at hindsight some of the incorporated ideas are worth a second look. When finally everything was in place at the beginning of the summer of 1990 the deadline for the proceedings of the 1990 DARPA Image Understanding Workshop just passed. Azriel nevertheless succeeded to squeeze in a short four page note (Mintz et al., 1990b), while the full version became a technical report in December (Mintz et al., 1990a). It was submitted to IEEE PAMI, lingered for about three years on the desk of an Associate Editor and finally faded out through attrition. The CBD paper was the only publication on robust methods from CfAR which did not get a wide exposure.

A much more pleasant publishing experience was the generalized minimum volume ellipsoid (GMVE) method for feature space analysis. Jean-Michel Jolion from the Claude Bernard University, Villeurbanne, France, visited twice CfAR in the late 80's. During the first visit, on a NATO fellowship between September 1987 to September 1988, he became the most active member of the pyramid group and later wrote with Azriel a book on image pyramids (Jolion and Rosenfeld, 1994). During his second visit, in the summer of 1989, Jean-Michel became familiar with robust estimation and we started to work on a robust clustering method.

The GMVE approach was very simple. Use the minimum volume ellipsoid as a computational module to find the currently most significant cluster in the feature space. Once delineated the cluster is removed and the procedure repeated till all the data points were classified. Note that this peeling-off approach did not require a priori knowledge of the number of clusters in contrast with most traditional techniques. As long as the feature space was not too complicated (contained only a few clusters), and more importantly a decomposition through the elliptical tiles imposed by the MVE did not introduce too severe artifacts, the method had an excellent performance.

The GMVE was used in several robust clustering applications: range image segmentation of man made objects, analysis of the Hough accumulator, histogram based gray level image segmentation. Working remotely by E-mail in those pre-Internet days, we finished a paper in the Fall of 1990 and submitted it to IEEE PAMI in December. The reviews were unusually positive and the paper was in print in a record time by August 1991 (Jolion et al., 1991).

In spite of its success we knew that GMVE was not the ultimate solution for feature space analysis. For example, the complexity of a color space derived from an outdoor image was beyond the capacity of GMVE to produce a reliable decomposition, and thus to segment the image. Such complex spaces can be handled only with a fully nonparametric methods, and it took several more years to develop one (Comaniciu and Meer, 1999).

Interestingly, the basic computational module of the nonparametric robust clustering technique, the mean shift procedure, is again and old largely forgotten pattern recognition method proposed more than 25 years ago (Fukunaga and Hostetler, 1975). The mean shift is based on the observation that the vector of the mean of the data points in a window is proportional to the locally estimated gradient of the density of these points. Thus, recursively moving the window along the mean vectors will lead to a region of maximum density, i.e., to a local mode. A cluster is then delineated by the basin of attractions of the mode and its shape its unconstrained.

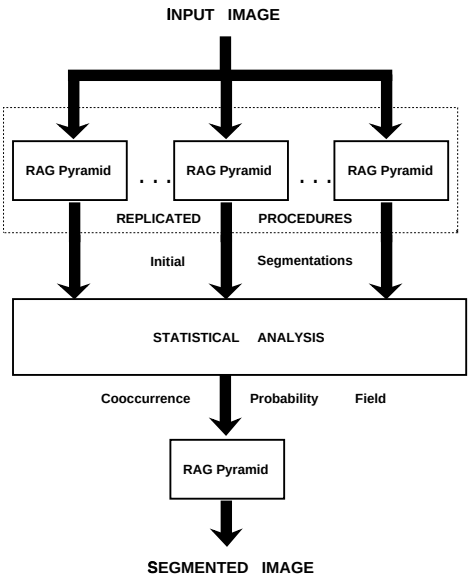
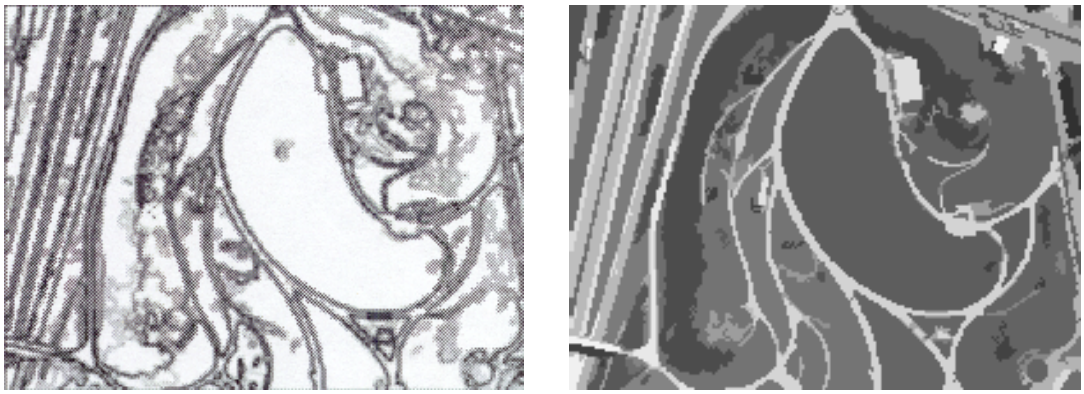


Figure 1.10 Information flow in a consensus based image segmentation algorithm.

There is a second connection between the multiresolution image analysis, as performed by the region adjacency graph (RAG) pyramids and modern statistics. The probabilistic component of the RAG pyramid based image segmentation implies that each time the algorithm is run the result will be slightly different. The differences will be more significant in the neighborhoods where the assumed homogeneity model (piecewise constancy) is less valid. This important information can be extracted by pixelwise comparison of the labeled images, for example, by defining the co-occurrence probabilities for pairs of adjacent pixels. In Figure 1.10 information flow of the approach is shown. For visualization the eight-dimensional vector of a pixel's co-occurrence probabilities has to be transformed into a scalar. The scalar field is then represented as a gray level image, called the consensus image. In Figure 1.11a an example is shown for the input image in Figure 1.1a. Note the richness of the extracted information, which can be exploited for a satisfactory image segmentation (Figure 1.11b). The segmentation algorithm is described in (Cho and Meer, 1997).

The above sketched approach is more general than just a method to improve image segmentation. First the relation between the input data and the algorithms applied to it is slightly perturbed. If after the perturbation the output remains an unbiased estimate of the desired (true) result, by combining the perturbed outputs a more reliable final outcome can be obtained. When we realized this, quickly a short paper was submitted to the next available workshop



(a) (b)

Figure 1.11 An example of a consensus based image segmentation. (a) A consensus image obtained from 20 initial segmentations for the image in Figure 1.1a. (b) Segmented image.

describing the basic principles of what we called the consensus paradigm (Meer et al., 1990a). Beside image segmentation, the paradigm was also employed for edge detection (Mintz, 1991).

The consensus paradigm was a rather ad-hoc approach to improve the quality of the output of a complex vision task. However, it turned out that the underlying principle is very similar to that of the bootstrap methodology introduced in statistics in 1979 by Efron. Bootstrap is a technique to obtain statistical measures from the available *single* data sample. It has solid theoretical foundations and is slowly becoming a standard statistical tool. The book (Efron and Tibshirani, 1993) provides an excellent introduction.

The idea behind bootstrap is to consider the available data as an empirical distribution and generate "new" data, i.e., bootstrap samples from it by sampling with replacement. The output of interest is computed for each bootstrap sample and thus from the single input a distribution of outputs is obtained. (Note the similarity with the approach shown in Figure 1.10.) This distribution can then be used to obtain the statistical measures of interest, like the covariance matrix of the output corresponding to the original data. Since in computer vision often analytical evaluation of the reliability of an estimate at the output of a task with several computational steps is not feasible, bootstrap can be a valuable tool to achieve this goal. See (Matei and Meer, 1999) for an example where bootstrap generated covariances for estimated 3D locations were used in rigid motion estimation of a stereo head.

Today the importance of using the proper statistical formulation when solving an image understanding problem is widely recognized in the vision commu-

nity. A lot of progress has been made in the last decade. To see this it is enough to compare the first attempts of applying robust estimators to vision problems (most of them mentioned in this paper), and the papers in the recent special issue on "Robust Statistical Techniques in Image Understanding" which appeared in April 2000 in the journal *Computer Vision and Image Understanding*. While many of the problems addressed are similar, the proposed tools became much more sophisticated. They are no longer straightforward "imports" from statistics but are trying to suit the peculiarities of visual data. This hardly gained awareness of what is needed to have reliable vision algorithms, raised more open questions about optimal design of image understanding tools than we had ten years ago. The availability of a computational power nobody could have predicted in 1990, the ubiquity of visual information in the Internet era, are just further increasing the pressure on us to finally deliver, what in the 1960's was assumed to be a trivial task: a general, autonomous vision system.

4. INSTEAD OF CONCLUSIONS

There is a subtle message in the history of robust estimation in computer vision. It turned out that the best methods: Hough transform, RANSAC, mean shift were all developed independently from statistics, and except of Hough transform had to be rediscovered. The rigor of statistics is certainly a good thing when defining problems, however, the role of intuition should not be underestimated either. One cannot stop asking the question, how many hidden treasures are still out there in the forgotten literature of the founding fathers?

The main goal of this narrative was to give a (possibly subjective) glimpse of a narrow slice of activities in the Computer Vision Laboratory in the late 80's. The continuous parade of visitors, the open exchange of ideas, the quick reactions to assure priority for important result characterized the entire lab. It was a wonderful place to be part of, and you always felt as being in one of the focal points of the computer vision community.

Today the community is much more distributed and research is often solely application driven. Nevertheless, many of us who had the privilege of passing through CfAR in those days are trying to preserve in our own laboratories that special atmosphere of intellectual freedom which characterized CfAR. Without a doubt, this is also part of Azriel Rosenfeld's legacy to the vision community.

References

Baronti, S., Casini, A., Lotti, F., Favaro, L., and Roberto, V. (1990). Variable pyramid strucutres for image segmentation. *Computer Vision, Graphics, and Image Proc.*, 49:346–356.

Basseville, M. and Benveniste, A. (1983). Design and comparative study of some sequential jump detection algorithms for digital signals. *IEEE Trans. Acoust., Speech, Signal Proc.*, 31:521–535.

Besl, P. J., Birch, J. B., and Watson, L. T. (1988). Robust window operators. In *Proceedings of the 2nd International Conference on Computer Vision*, pages 591–600, Tampa, FL.

Bister, M., Cornelis, J., and Rosenfeld, A. (1990). A critical view of pyramid segmentation algorithms. *Pattern Recognition Letters*, 11:605–617.

Bolles, R. C. and Fischler, M. A. (1981). A RANSAC-based approach to model fitting and its application to finding cylinders in range data. In *Proceedings of the 6th International Joint Conference on Artificial Intelligence*, pages 637–643, Vancouver, Canada.

Burt, P. J., Hong, T. H., and Rosenfeld, A. (1981). Segmentation and estimation of region properties through co-operative hierarchical computations. *IEEE Trans. Systems, Man, and Cybern.*, 11:802–809.

Cho, K. and Meer, P. (1997). Image segmentation from consensus information. *Computer Vision and Image Understanding*, 68:72–89.

Chow, G. C. (1960). Tests of equality between sets of coefficients in two linear regressions. *Econometrica*, 28:591–605.

Comaniciu, D. and Meer, P. (1999). Mean shift analysis and applications. In *Proceedings 7th International Conference on Computer Vision*, Kerkyra, Greece, pages 1197–1203.

Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. Chapman & Hall.

Fischler, M. A. and Bolles, R. C. (1981). Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Comm. ACM*, 24:381–395.

Fukunaga, K. and Hostetler, L. D. (1975). The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Trans. Information Theory*, 21:32–40.

Hansen, R. R. and Chellappa, R. (1988). Two-dimensional robust spectrum estimation. *IEEE Trans. Acoust., Speech, Signal Proc.*, 36:1051–1066.

Haralick, R. M. and Joo, H. (1988). 2D-3D pose estimation. In *Proceedings of the 9th International Conference on Pattern Recognition*, pages 385–391, Rome, Italy.

Hong, T. H., Shneier, M. O., Hartley, R. L., and Rosenfeld, A. (1983). Using pyramids to detect good continuation. *IEEE Trans. Systems, Man, and Cybern.*, 13:631–635.

Jolion, J. M., Meer, P., and Bataouche, S. (1991). Robust clustering with applications in computer vision. *IEEE Trans. Pattern Anal. Machine Intell.*, 13:791–802.

Jolion, J. M. and Rosenfeld, A. (1994). *A pyramid framework for early vision*. Kluwer.

Kashyap, R. L. and Eom, K. B. (1988). Robust image models and their applications. In *Advances in Electronics and Electron Physics*, volume 70, pages 79–157. Academic Press.

Kim, D. Y., Kim, J. J., Meer, P., Mintz, D., and Rosenfeld, A. (1989). Robust computer vision: A least median of squares based approach. In *Proceedings 1989 DARPA Image Understanding Workshop*, pages 1117–1134, Palo Alto, CA.

Kiryati, N. and Bruckstein, A. (1992). What's an a set of points? *IEEE Trans. Pattern Anal. Machine Intell.*, 14:496–500.

Kropatsch, W. G. (1995). Building irregular pyramids by dual-graph contraction. *IEE Proc.-Vis. Image Signal Process.*, 142:366–374.

Li, G. (1985). Robust regression. In Hoaglin, D. C., Mosteller, F., and Tukey, J. W., editors, *Exploring Data Tables, Trends, and Shapes*, pages 281–343. Wiley.

Matei, B. and Meer, P. (1999). Optimal rigid motion estimation and performance evaluation with bootstrap. In *Computer Vision and Pattern Recognition Conference*, pages 339–345, Fort Collins, CO.

Mathieu, C. E. and Magnin, I. E. (1996). On the choice of the first level on graph pyramids. *Journal of Mathematical Imaging and Vision*, 6:85–96.

Meer, P. (1989). Stochastic image pyramids. *Computer Vision, Graphics, and Image Processing*, 45:269–294.

Meer, P., Mintz, D., Kim, D. Y., and Rosenfeld, A. (1991). Robust regression methods in computer vision: A review. *International Journal of Computer Vision*, 6:59–70.

Meer, P., Mintz, D., Montanvert, A., and Rosenfeld, A. (1990a). Consensus vision. In *Proceedings AAAI-90 Workshop on Qualitative Vision*, pages 111–115, Boston, Mass.

Meer, P., Mintz, D., and Rosenfeld, A. (1990b). Least median of squares based robust analysis of image structure. In *Proceedings 1990 DARPA Image Understanding Workshop*, pages 231–254, Pittsburgh, PA.

Meer, P., Sher, C. A., and Rosenfeld, A. (1990c). The chain pyramid: Hierarchical contour processing. *IEEE Trans. Pattern Anal. Machine Intell.*, 12:363–376.

Mintz, D. (1991). Robust consensus based edge detection. Technical Report CAR-TR-546, Center for Automation Research. University of Maryland, College Park.

Mintz, D., Meer, P., and Rosenfeld, A. (1990a). Consensus by decomposition: A paradigm for fast high breakdown point robust estimation. Technical Report CAR-TR-525, Center for Automation Research. University of Maryland, College Park.

Mintz, D., Meer, P., and Rosenfeld, A. (1990b). A fast, high breakdown point robust estimator for computer vision applications. In *Proceedings 1990 DARPA Image Understanding Workshop*, pages 255–258, Pittsburgh, PA.

Montanvert, A., Meer, P., and Rosenfeld, A. (1991). Hierarchical image analysis using irregular tessellations. *IEEE Trans. Pattern Anal. Machine Intell.*, 13:307–316.

Peleg, S., Federbusch, O., and Hummel, R. A. (1986). Custom-made pyramids. In Uhr, L., editor, *Parallel Computer Vision*, pages 125–146. Academic Press.

Rosenfeld, A., editor (1984). *Multiresolution Image Processing and Analysis*. Springer.

Rosenfeld, A. (1986a). Pyramid algorithms for perceptual organization. *Behavior Research Methods, Instruments, & Computers*, 18:595–600.

Rosenfeld, A. (1986b). Some pyramid techniques for image segmentation. Technical Report CAR-TR-203, Center for Automation Research, University of Maryland at College Park.

Rousseeuw, P. and Leroy, A. (1987). *Robust regression and outlier detection*. John Wiley & Sons.

Sher, C. A. and Rosenfeld, A. (1990). A pyramid programming environment on the connection machine. *Pattern Recognition Letters*, 11:241–245.

Shneier, M. O. (1982). Extracting linear features from images using pyramids. *IEEE Trans. Systems, Man, and Cybern.*, 12:569–572.

Spann, M., Horne, C., and Du Buf, J. M. H. (1989). The detection of thin structures in images. *Pattern Recognition Letters*, 10:175–179.

Stewart, C. V. (1997). Bias in robust estimation caused by discontinuities and multiple structures. *IEEE Trans. Pattern Anal. Machine Intell.*, 19:818–833.

Tirumalai, A. and Schunk, B. G. (1988). Robust surface approximation using least median of squares. Technical Report CSE-TR-13-89, Artificial Intelligence Laboratory. University of Michigan, Ann Arbor.

Ziavras, S. G. and Meer, P. (1994). Adaptive multiresolution structures for image processing on parallel computers. *Journal of Parallel and Distributed Computing*, 23:475–483.