Elimination of Outliers from 2-D Point Sets Using the Helmholtz Principle

Demetrios P. Gerogiannis, *Member, IEEE*, Christophoros Nikou, *Senior Member, IEEE*, and Aristidis Likas, *Senior Member, IEEE*

***Abstract*—A method for modeling and removing outliers from 2-D sets of scattered points is presented. The method relies on a principle due to Helmholtz stating that every large deviation from uniform noise should be perceptible, provided that the deviation is generated by an *a contrario* model of geometric structures. By assuming local linearity, we first employ a robust algorithm to model the local manifold of the corrupted data by local line segments. Our rationale is that long line segments should not be expected in a noisy set of points. This assumption leads to the modeling of the lengths of the line segments by a Pareto distribution, which is the adopted *a contrario* model for the observations. The model is successfully evaluated on two problems in computer vision: shape recovery and linear regression.**

***Index Terms*—Linear regression, outlier modeling, point cloud, shape detection.**

I. Introduction

THE modeling and removal of outliers from a set of points has been an active research topic for many decades in image processing and computer vision and a variety of algorithms have been proposed [2]. They may be as simple as the median filter to be more elaborated which are based on random sampling, like RANSAC [10] or probabilistic models [15].

The Gaussian assumption for data generation has been widely adopted but it is appropriate only for sparse outlier distributions. In general, it involves the comparison of Euclidean distances between points with the mean of the distribution expanded by a number of standard deviation [13]. Kernel density estimators-based methods provide a probabilistic approach to determine if a point belongs to the uncorrupted set and are inherently related to clustering or classification techniques that separate pure data from outliers [14], [7].

The number of neighbors of a point is a key issue in characterizing it as outlier [8]. The main hypothesis is that pure data are more densely populated than outlying points and many algorithms have been designed based on this idea. The adopted strategy consists in defining a neighborhood for each point, determining a feature that characterizes the neighborhood and rejecting points with features having a value smaller than a threshold. In [18], the number of common neighbors is defined as a similarity index between points and points with neighborhood size smaller than a threshold are rejected as outliers. An octree is used in [19] to cluster points and an implicit quadric is fit to them to smooth out outliers.

Inspired by the geometric Gestalt theory, which addresses the answer to the fundamental problem in computer vision: "How to arrive at global percepts from the local, atomic information contained in an image?" [5], Desolneux *et al.* proposed methods for detecting geometric structures [3] and edges [4] in images by a parameter free method based on the Helmholtz principle [6]. The principle states that an observed geometric structure is perceptually meaningful if its number of occurrences is very small in a random situation. In this context, geometric structures are characterized as large deviations from randomness. The principle is accompanied by an *a contrario* assumption against which structures are detected.

In this paper, we propose an algorithm for outlier elimination and structure extraction from 2-D point clouds based on the Helmholtz principle. The main difference with the methods in [3], [4] is that the input to the algorithm is not an image whose pixels lay on a regular grid but a set of scattered points irregularly distributed in space. To overcome this limitation, at first, the point set is approximated by a locally linear manifold consisting of a set of line segments [11]. We show that the lengths of the line segments follow a Pareto distribution which is our *a contrario* model.

In the remainder of the paper, the Helmholtz principle is presented in Section II, the extraction of meaningful line segments and the outlier modeling are described in Section III and numerical experiments are discussed in Section IV.

II. The Helmholtz principle

The Helmholtz principle is a general hypothesis of the Gestalt theory [5] interpreting how the human perception works. Intuitively, it states that if we take into consideration randomness as the normal case for our observations then meaningful features and interesting events should not be expected. Consequently, if they are observed they should appear with a small probability. Moreover, the small probability of observing an event is not a factor to consider it as meaningful (or true observation not generated by noise). Take as an example the toss of an unbiased coin. The probability of getting either a head (H) or a tail (T) is 1/2. If we toss the coin successively N times then the probability of observing any of the possible sequences of H and T

Manuscript received February 12, 2015; revised April 02, 2015; accepted April 03, 2015. Date of publication April 07, 2015; date of current version April 15, 2015. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Edmund Lam.

The authors are with the Department of Computer Science and Engineering, University of Ioannina, Ioannina GR 451 10, Greece (e-mail: dgerogia@cs.uoi.gr; cnikou@cs.uoi.gr; arly@cs.uoi.gr).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LSP.2015.2420714

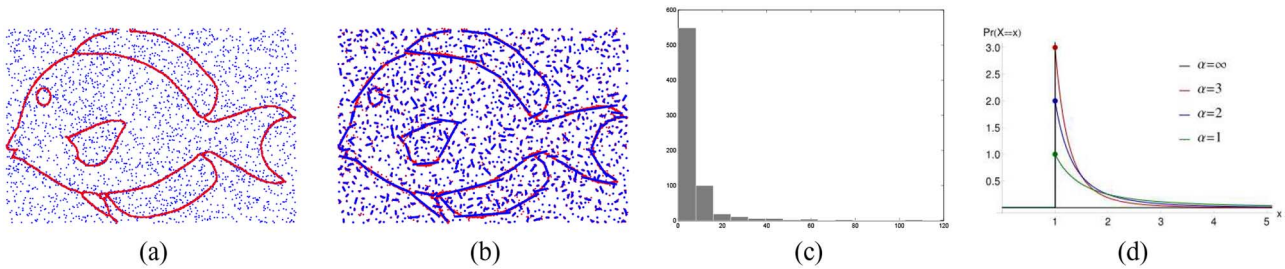

Fig. 1. (a) A a set of points (in red color) degraded by equal in number outliers (in blue color). (b) A line fitting example of the points in (a). (c) The distribution of the lengths of the line segments approximating the point set of (a) using a line segment detection algorithm [11]. The horizontal axis represents the lengths and the vertical axis represents the corresponding frequencies. (d) The Pareto distribution for various values of the parameter a with $b = 1$.

is $(1/2)^N$, which is a decreasing function of N and approaches zero as $N \to \infty$. More specifically, the following sequences:

$$S_1 = \underbrace{THTHHTT \dots HT}_{N \text{ times}}, S_2 = \underbrace{HHHHHHH \dots H}_{N \text{ times}}$$

have equal probabilities of appearance. However, S_2 is not expected to appear for an unbiased coin. Therefore, the low probability of an event may not characterize it as a deviation from randomness, as its probability may not truly model the randomness of an event.

Using the same sequences S_1 and S_2, we may define another pair of random variables n_H and n_T modeling the number of H and T present in a sequence. Since the coin is unbiased, the expectations of both variables is $N/2$. Although this is confirmed in S_1, in sequence S_2 the observed values for n_H and n_T is N and 0 largely deviating from the expected values.

The above observations lead to the conclusion that the small probability of an event may not be an accurate indication that this event is meaningful and we need to take into consideration that the model we use to validate an event describes the randomness of all possible observations. Turning back to the last example of the coin toss, randomness was modeled only by counting the number of H and T in a sequence and not by the probability of a sequence to appear. Taking both issues into account yields the complete model used to describe randomness which is called *a contrario* model.

III. Meaningful Line Segments and Outliers

Let $X = \{\mathbf{x}_i\}_{i=1,\dots,N}$ be a set of observed 2-D points including both data points and outliers (Fig. 1(a)).

In order to eliminate the outliers, we compute at first an approximation of the point set by line segments (e.g. [11], [12]). In the example of Fig. 1(a), the large number of outliers will provide a large number of line segments with relatively small lengths (due to noise) and a smaller number of line segments with larger lengths (due to both the uncorrupted data and the noise around them), as shown in Fig. 1(b). This distribution of the lengths of the line segments, shown in Fig. 1(c) leads to consider an *a contrario* probabilistic model of the lengths described by a Pareto distribution [1]:

$$\text{Pareto}(x; a, b) = \begin{cases} \frac{ab^a}{x^{a+1}}, & x \geq b \\ 0, & x < b \end{cases} \quad (1)$$

where $b > 0$ and a is a parameter controlling the slope of the curve (Fig. 1(c)). Herein, the length of the segment is considered in terms of the number of the points contributed to its computation. The line segment detection algorithm provides line segments with uniformly distributed points, e.g. [11], [12]. Therefore, the length of a segment is equivalent to the number of points belonging to it.

The purpose of the *a contrario* model is to describe the randomness of the data. However, it might be possible that outliers are organized in such a way that they generate short line segments that are not part of the desired structure. The Pareto distribution computes the probability that a segment of a given length appears in the observations. In an analogy to the coin toss example, this event may be expressed by the probability of getting H or T (with more possible outcomes, which are the lengths of line segments). By expanding our initial intuition regarding the rareness of the observation, it is possible that segments due to outliers would be isolated, as the intrinsic feature of noise is to be structureless. Therefore, in order to set up the *a contrario* model, the neighborhood of a line segment should be defined to account for isolated structures.

Each line segment has a starting and an ending point. The neighborhood $\mathcal{N}(\beta)$ of a segment β is defined as the set of all those segments β_j whose starting/ending points are located at a distance less than a threshold to the starting/ending points of β:

$$\mathcal{N}(\beta) = \{\beta_j : |\beta^k - \beta_j^l| \leq T, k, l \in \{s, t\}\}, \quad (2)$$

where the superscripts $\{s, t\}$ indicate the starting or the ending point of a segment. The neighborhood can be iteratively expanded to take into account the neighbors of neighbors up to a fixed depth.

Therefore, the *a contrario* model is based on the assumption that a line segment is more probable to be a true observation if its neighboring segments have large lengths. This may be expressed by the likelihood:

$$\mathcal{L}(\beta) = \prod_{\beta_j \in \mathcal{N}(\beta)} \text{Pareto}(\beta_j; a, b). \quad (3)$$

Consequently, if $\mathcal{L}(\beta) < \epsilon$ we consider the line segments to be a true observation. The threshold is automatically determined as $\epsilon = 10^{-a}/D$, where D is the maximum depth of the neighborhood expansion. It may be observed that the value of ϵ is independent from the data.

The procedure is presented in Algorithm 1.

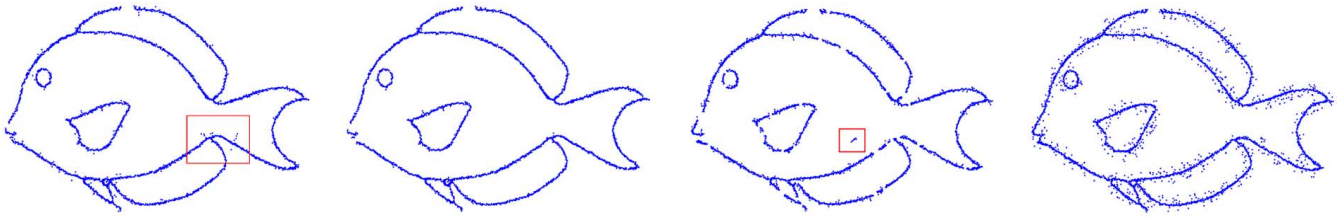

Fig. 2. Outlier elimination from the data set of Fig. 1(a) by (a) the first and (b) the last iterations of the proposed method, (c) Xianchao et al. [18], (d) DBScan [9]. The red boxes highlight representative false points provided by the methods.

Algorithm 1 Outlier elimination based on the Helmholtz principle.

input: A set of points X, the depth of expansion D.

output: A set of points Y.

while convergence is not reached **do**

Summarize X by line segments. Let B_i be the points contributing to segment β_i, for $i = 1 \dots N$.

$Y = \emptyset$.

for $i = 1 \dots N$ **do**

if $\mathcal{L}(\beta_i) \leq 10^{-a}/D$ **then**

$Y = Y \cup B_i$.

end if

end for

end while

IV. Experimental Results

To investigate the efficiency of the proposed method for outlier elimination, we used the Gatorbait database [16]. Degradation of the data set was artificially performed in the following way. For each point of the original data set, an outlier was generated by multiplying the coordinates of that point with a uniformly distributed random number in the interval (0,1]. The number of outliers added was set equal to the number of pure data points. Moreover, the pure data were degraded by zero-mean additive Gaussian noise with an appropriate standard deviation in order to obtain a signal to noise ratio (SNR) of 55 dB (e.g. Fig. 1(a)). The algorithm was applied to 50 different realizations of outliers, in order to obtain more accurate results that are not biased to a specific configuration.

We conducted comparisons with a density-based method (DBScan [9]) and the algorithm of Xianchao *et al.* [18]. Let us also note that other established methods, such as the algorithm in [15], were also considered but they failed to provide an acceptable result in our framework of highly corrupted point sets. Finally, we also show the results of the simple, but in many cases powerful, median filter for image denoising to highlight the order of magnitude of the obtained accuracy with respect to a well known baseline.

TABLE I
Statistics on the Hausdorff Distance (4) on the 38 Shapes of the GatorBait100 data set [16]

Method	mean	std	median	min	max
Proposed method (a = 2)	**6.12**	**1.3**	**5.8**	**4.2**	**10.3**
Proposed method (a = 3)	**6.07**	**1.3**	**5.8**	**4.1**	**10.3**
Proposed method (a = 4)	**6.05**	**1.4**	**5.6**	**4.0**	**10.3**
Proposed method (a = 5)	**5.99**	**1.3**	**5.7**	**4.1**	**10.3**
DBScan [9]	12.62	3.1	11.2	10.5	23.1
Xianchao et al. [18]	84.59	19.6	83.0	51.4	129.7
Median Filter	208.17	17.0	208.4	174.9	243.8

To evaluate the results provided by the different algorithms we employed the Hausdorff distance between two sets of points X and Y:

$$d_{\mathcal{H}}(X,Y) = \max_{\mathbf{x}\in X} \min_{\mathbf{y}\in Y} \{|\mathbf{x}-\mathbf{y}|\}, \tag{4}$$

where X is the original set of points (the ground truth) and Y is the computed set of points after outliers removal.

Table I summarizes the performance of the compared methods, with the results of the proposed being marked in bold. As it may be seen, our method may successfully recover the initial shape. Its maximum distance 10.3 pixels, although smaller than the other algorithms, is due to the fact that, in a few cases, parts of the pure data were pruned because the outliers were close to them. Moreover, we examined the sensitivity of our method to parameter a of the Pareto distribution by applying the algorithm using a variety of values for this parameter, namely $a = \{2, 3, 4, 5\}$. As it may be observed, the method is consistent and its performance does not depend on this parameter. Larger values of a may not be employed as the numerator in the Pareto distribution (1) increases beyond computer accuracy. As b is the mode of the distribution, we have set $b = 2$ in all of the experiments relying on the fact that we search line segments and any two points define a line segment. Larger values of b, favor larger longer segments, and may account for the elimination of some details of the initial data that are modeled by shorter line segments. This relatively low value for b is not in favor of our algorithm, as the model accounts for less populated line segments which generally are due to noise. However, the results showed the robustness of the proposed approach. Furthermore, it is worth noting that DBScan [9] needs tedious parameter tuning (performed here by trial and error) and the method in [18] did not detect many outliers laying near the shape contour. Representative results are shown in Fig. 2. Xianchao *et al.* [18] preserves much of the initial information, but fails to remove outliers near the contour. The discontinuities provided by our method are due to

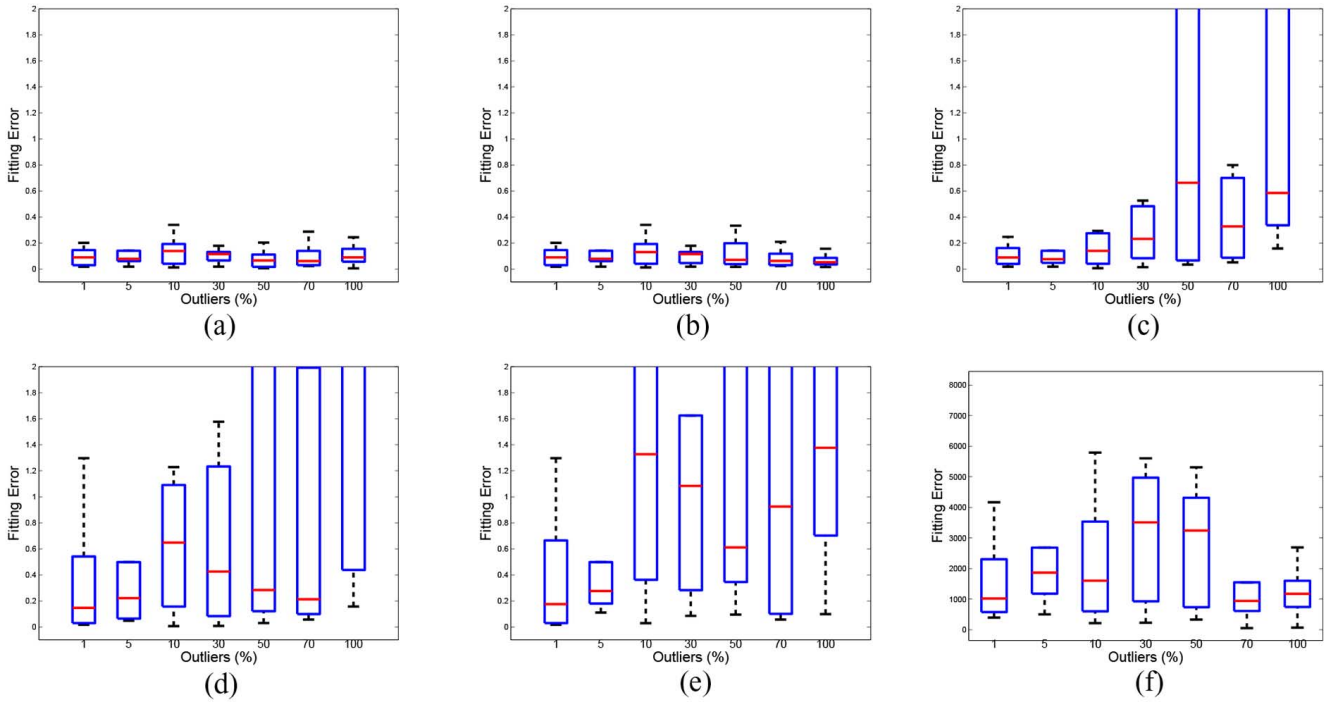

Fig. 3. Boxplots of the line fitting errors for the compared methods. Notice the different scales at the abscissas. (a) Proposed method (first iteration). (b) Proposed method (last iteration). (c) Xianchao et al. [18]. (d) DBScan [9]. (e) RANSAC [10]. (f) Dumouchel et al. [7].

the fact that the line segment fitting algorithm used many short segments to approximate the local manifold. Thus, the proposed framework rejected the corresponding points as outliers.

A second set of experiments addresses the problem of outlier elimination for line fitting. Following the same principles as in the previous experiments, a set of 500 colinear points were corrupted by outliers and Gaussian noise. Various experiments were conducted with an increasing number of outliers at each configuration. In the more challenging setup, the number of outliers was equal to the number of points. Each experiment was repeated 50 times and statistics on the fitting error, in terms of Euclidean distance between the estimated and the true parameters of the lines were computed. The performances of the compared methods are shown in Fig. 3. For a more meaningful evaluation, we have also compared our method with two robust algorithms, namely RANSAC [10] and the robust regression method proposed in [7]. As it may be seen, our algorithm outperforms both of these methods which are established in the computer vision literature. Please notice the different scales in the abscissas in the graphs in Fig. 3 which clearly show the accuracy of the proposed algorithm as its maximum error, even in the more challenging scenario is less than one coordinate unit. On the other hand, only RANSAC is relatively competitive but its fitting errors are larger.

A final experimental configuration investigated the dependence of the proposed framework to the selected line segment detection algorithm. To that end, the test image of Fig. 4(a) was used, where the circle is considered as an outlying shape and the other shapes consisting of line segments are the inliers. The assumption is that the circular shape needs a large number of short line segments to be approximated. We have employed two line detection algorithms for the corresponding step of Algorithm 1: DSaM [11] and polygon approximation (PA) [12]. The parameters for the Helmholtz principle were $a = 5$ and $D = 3$. Also, $b = |\text{mean}(L) - \text{std}(L)|$, where $L = \{L_i\}, i = 1 \dots K$, with L_i being the number of points contributing to the computation of the i-th line segment and K is the total number of line segments. Figs. 4(b)–(c) represent the output of the outlier elimination method based on DSaM [11] and PA [12] respectively. It may be observed that both methods provide similar results, which confirms that the proposed framework is consistent independently of the line segment detection algorithm selected. Moreover, the Helmholtz principle enables the determination of a more elaborated criterion that can trim outliers at a higher level (e.g. remove circles in a set with linear structures). This cannot be achieved with a standard density method, e.g. Xianchao et. al [18], as can be seen in Fig. 4(d), where a significant part of the linear structures was also removed.

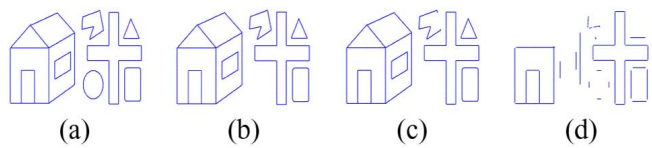

Fig. 4. (a) A test image and the outlier removal result based on (b) DSaM [11], (c) polygon approximation PA [12] and (d) the method in [18].

V. Conclusions

A method for removing outliers from unordered point clouds has been presented. The method relies on the Helmholtz principle, which states that in a random model meaningful observations should not be expected. We showed that approximating the local manifold of the scattered points yields an efficient modeling of the outliers. Comparisons with established algorithms for outlier elimination and robust regression highlighted that the proposed method consists an efficient framework.

REFERENCES

[1] B. C. Arnold, *Pareto Distributions*, ser. Statistical Distributions in Scientific Work Series. London, U.K.: Chapman & Hall, 1983.

[2] C. C. Aggarwal, *Outlier Analysis*. Berlin, Germany: Springer, 2013.

[3] A. Desolneux, L. Moisan, and J. M. Morel, "Meaningful allignments," *Int. J. Comput. Vis.*, vol. 40, pp. 7–23, 2000.

[4] A. Desolneux, L. Moisan, and J. M. Morel, "Edge detection by helmholtz principle," *J. Math. Imag. Vis.*, vol. 14, pp. 271–284, 2001.

[5] A. Desolneux, L. Moisan, and J. M. Morel, "Gestalt theory, and computer vision," in *Seeing, Thinking and Knowing*. Norwell, MA, USA: Kluwer, 2004, vol. 38, pp. 71–101.

[6] A. Desolneux, L. Moisan, and J. M. Morel, *From Gestalt Theory to Image Analysis: a Probabilistic Approach*. Berlin, Germany: Springer, 2008.

[7] W. Dumouchel and F. O'Brien, "Integrating a robust option into a multiple regression computing environment," in *Computing and Graphics in Statistics*. New York, NY, USA: Springer-Verlag, 1991, pp. 41–48.

[8] L. Ertöz, M. Steinbach, and V. Kumar, "Finding clusters of different sizes, shapes, and densities in noisy, high dimensional data," in *SIAM Int. Conf. Data Mining*, 2003.

[9] M. Ester, H.-P. Kriegel, S. Jórg, and X. Xiaowei, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *International Conference on Knowledge Discovery and Data Mining*. Cambridge, MA, USA: AAAI Press, 1996, pp. 226–231.

[10] M. A. Fischler and R. C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[11] D. Gerogiannis, C. Nikou, and A. Likas, "Modeling sets of unordered points using highly eccentric ellipses," *EURASIP J. Adv. Signal Process.*, vol. 2014, no. 1, p. 11, 2014.

[12] U. Ramer, "An iterative procedure for the polygonal approximation of plane curves," *Comput. Graph. Image Process.*, pp. 244–256, 1972.

[13] R. B. Rusu and S. Cousins, "3D is here: Point cloud library (PCL)," in *IEEE International Conference on Robotics and Automation*, Shanghai, China, May 9–13, 2011.

[14] O. Schall, A. Belyaev, and H.-P. Seidel, "Robust filtering of noisy scattered point data," in *Eurographics Symp. Point-Based Graphics*, 2005, pp. 71–77.

[15] R. Thompson, "A note on restricted maximum likelihood estimation with an alternative outlier model," *J. Roy. Statist. Soc.*, vol. 47, no. 1, pp. 53–55, 1985.

[16] Univ. Florida, Gatorbait 100 [Online]. Available: http://www.cise.ufl.edu/anand/publications.html, Jan. 2014

[17] T. Weyrich, M. Pauly, R. Keiser, S. Heinzle1, S. Scandella, and M. Gross, "Post-processing of scanned 3D surface data," in *Eurographics Symp. Point-Based Graphics*, 2004, pp. 85–94.

[18] W. Xiaochao, L. Xiuping, and Q. Hong, "Robust surface consolidation of scanned thick point clouds," in *Int. Conf. Computer-Aided Design and Computer Graphics (CAD/Graphics)*, 2013, pp. 38–43.

[19] H. Xie, K. T. McDonnell, and Q. Hong, "Surface reconstruction of noisy and defective data sets," *IEEE Visualization*, pp. 259–266, 2004.