A new approach for removing point cloud outliers using box plot

Hafsa Benallal, Youssef Mourchid, Ilyass Abouelaziz, Ayman Alfalou, Hamid Tairi, Jamal Riffi, Mohammed El Hassouni

A new approach for removing point cloud outliers using box plot

Conference Paper · May 2022

DOI: 10.1117/12.2618842

7 authors, including:

Hafsa Benallal
Sidi Mohamed Ben Abdellah University
2 PUBLICATIONS **1** CITATION
SEE PROFILE

Youssef Mourchid
CESI
22 PUBLICATIONS **92** CITATIONS
SEE PROFILE

Ilyass Abouelaziz
Mohammed V University of Rabat
25 PUBLICATIONS **181** CITATIONS
SEE PROFILE

H. Tairi
Sidi Mohamed Ben Abdellah University
213 PUBLICATIONS **1,154** CITATIONS
SEE PROFILE

A New Approach for Removing Point Cloud Outliers using Box plot

Hafsa Benallal[a], Youssef Mourchid[b], Ilyass Abouelaziz[b], Ayman Al Falou[b], Hamid Tairi[a], Jamal Riffi[a], and Mohammed El Hassouni[c]

[a]LISAC Laboratory, Sidi Mohamed Ben Abdellah University, Fez, Morocco
[b]L@bISEN, LSL Team, Yncréa Ouest, Brest, France
[c]FLSH, Mohammed V University in Rabat, Rabat, Morocco

ABSTRACT

Cleaning a point cloud building is challenging issue, it is crucial for a better representation of the scan-to-BIM 3D model. During the scan, the point cloud is in generally influenced by several factors. The scanner can provide false data due to reflections on reflective surfaces like mirrors, windows, etc. The false points can form a whole bunch of disturbing data which is not easy to detect. In this work, we use a statistical method called box plot to clean the data from false points. This method is a developed method of reading histograms. We test the proposed method on private database containing four point cloud buildings specifically designed for building information modeling (BIM) application. The experimental results are satisfying and our method detect most of the false points in the database.

Keywords: Point cloud, Building information modeling, Outlier removal, box plot

1. INTRODUCTION

The pre-processing step will be explored in order to clean the point clouds to lighten the database. The idea is to develop a new model based on standard statistical approaches for this kind of processing. According to (Hawkins et al, 1980)[1] an outlier is a value located at the extreme value of a data series, the most commonly it can be generally the minority of observations in a dataset that have different trends than the majority of observations in the dataset, the detection of outliers is part of the data analysis and cleaning.[2] In some cases, they can interfere to our BIM project if we don't manage it well. Box plot are among the families of statistical tools widely used in exploratory data analysis (EDA). They aim to represent a distribution schematically and to summarize a variable in a simple and visual way, as well as to identify extreme values.[2] Stated "There are numerous definitions of outliers in the statistical and machine learning literatures". A box plot is a graph that gives us a good indication of the distribution of data. Although box plot can be compared to a histogram or a density plot, they have the advantage of taking up less space, which is useful for comparing distributions in datasets. Thus, the detection of outliers has received and continues to receive attention from the statistical literature. see for example.[3] The outliers in our work are in the form of either scattered points in space or a group of points that are due to the bad acquisition of the scanner. In this article, we propose another method of building box plot, the structure of the document is as follows: in section 1, we describe our recommendations. In Section 2, we compare our method with existing methods. Section 3 is dedicated to comparative results. In Section 4, we apply our methodology to some well-known datasets that cross outliers, where traditional methods seem unable to detect outliers.

2. RELATED WORK

Given the many fields of application that require the detection of outliers, the literature offers a large number of methods dedicated to it.[4] In general, the outlier detection technology can be divided into 7 main groups, each group is based on distribution, density, distance, depth, classification, clustering and Decomposition or projection of spectral data as shown in the figure 1. Among the most known methods, are those based on distribution also known as statistical based, these strategies point to show the dissemination of information with a known or obscure work, a parametric or non-parametric approach.[1] In this type of methods any outlier is considered as a

extremely deviates of the standard data point. Density-based this method assume that the density of observations close to outliers is significantly less than the density of observations close to or more certain expected values.[5] These techniques are sometimes preferred to distance-based methods when the dataset consists of subsets of different densities. Thresholds have been established to determine the vicinity of an observation and the limit beyond which a value is considered abnormal. Distance-based method it generally uses the distance between the observation and the dataset to assess whether the observation is an outlier.[6–9] Euclidean distance and Mahalanobis distance are often reported in these methods. Several works have shown that current estimates of the location and size of the dataset. The mean and variance, are strongly influenced by outliers and the threshold is similar to the threshold for density-based methods. In depth-based strategy the dataset is modeled by a set of convex envelopes whose purpose is to group data based on their proximity to the core of the dataset,[10] The convex envelope near the core of the dataset contains the most expected values, while the outer envelope distinguishes the outliers. The computation time becomes too long when the dataset dimension exceeds 3.[11] Classification-based methods can be supervised or unsupervised. In the first case a literacy dataset for which the position of outliers is known is used to construct a classifier that will distinguish outliers in a confirmation dataset. In the other case, the classifier attempts to model the data set without using a test or a confirmation dataset. Clustering-based these approaches seek to group observations into different subsets and consider outliers to be process residues. Observations that are not related to a subset.[7,12] Intuitively, the latest observations to be linked are more likely to be outliers. These methods are sometimes criticized because their output depends heavily on the choice of clustering algorithm. Moreover, since these approaches were primarily designed for the clustering of data, their potential for detecting outliers could be questioned.[13,14] Spectral decomposition-based Observations are projected into a new subspace to facilitate the detection of outliers.[15,16] This approach is preferred when the dataset has multiple dimensions. When other methods would not produce a result within a reasonable time.

These techniques provide a solution to the so-called curse of dimensionality. This problem is because increasing the number of dimensions of the data, the outliers will be increasingly rooted in the dataset and will be increasingly difficult to detect. The number of main components to choose for decomposition remains a challenge and is always defined by a threshold. Hybrid approaches have been proposed by several works with combining several of these techniques,[8,17] strategies based are distance, density and clustering are most often mentioned.

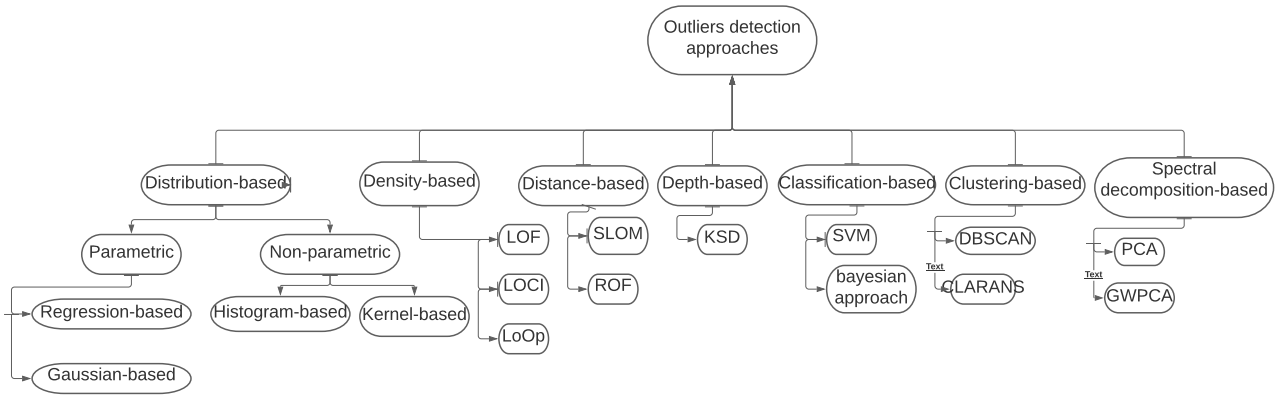


Figure 1. outliers detection methods

3. OUR METHOD

3.1 Flowchart for our method

Filtering process for plot box as showna bove 2, is a graphical representation of the general structure of a box plot, it describes the flow of data from the input data to the final equation, this information processing system describes the flow of set of logical operations passing by reordering set of data then choosing the median according to whether the number is even or odd then calculating quartile Q1 and Q3 to arrive at the end of interquartile range IQR as indicated in the equations 1, 2, 3, 4. The box plot is a graphical representation to visualize data with more detailed information so it is very useful when it comes to understanding the distribution of data. It shows outliers more clearly.

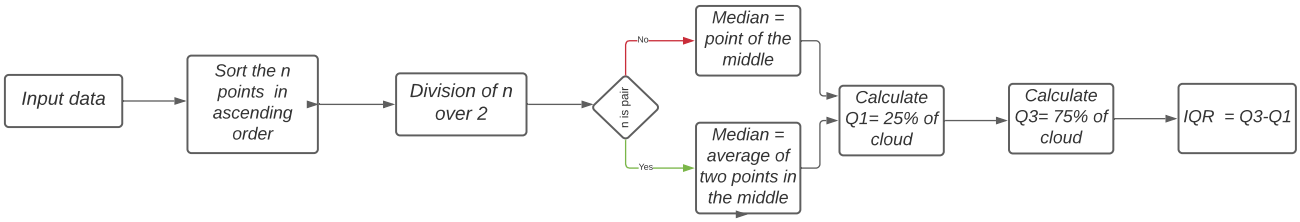


Figure 2. The box plot filtering process

Quartile is the central value of the data set that describes in statistics a division of the observations into four intervals defined on the basis of the data values and their comparison with the set of observations. The median is the value that separates the bottom half from the top half of a sample set. Intuitively, the median is thus the midpoint of the set 2. Then comes the Quartile Q1 is the intermediate number between the lowest number and the median of the data set and is not the "minimum" 1.

The Quartile Q3 is the middle number between the median and the highest value of the data set and not the "maximum" 3. The range between Q1 and Q3 called the interquartile range (IQR) calculated by Q3-Q1 4.

We can characterize outliers in dataset if it has values below than **Q1-1.5 * IQR** and above than **Q3 + 1.5 * IQR**, and all data point that is located outside the whiskers of the box plot.

The end of the lower whisker is the smallest value in the data that is greater than the lower limit Q1 - 1,5*(Q3-Q1). The end of the upper whisker is the maximum value in the data that is less than the high limit Q3 +1,5*(Q3-Q1).

Based on the cloud distribution and with a well chosen threshold from the associated axis histograms we will be able to combine the filtration along at least two axes in the figure **??**. For each component xi in Pi if value is between Q1 - threshold * IQR and Q1 + threshold * IQR then this value is considered as non outlier if not the point will be deleted.

3.2 Adjustment to the box plot

In statistics it is easier to visualize outliers and because of this statisticians were the first to detect an outlier, so the detection of outliers became a challenge. the plot box gives a simple representation by summarizing the main terms lower extreme, lower quartile, median, upper quartile and upper extreme. It aims first at the quantitative data which are in our case the point cloud of different building. The thresholds are calculated so that each column of data will have its own upper and lower quartile and median. for example for X axis the quartile is calculated as follows :

$$Lower\ quartile\ (Q1) = (N+1) * 0.25\,, \tag{1}$$

$$Middle\ quartile\ (Q2) = (N+1) * 0.50\,, \tag{2}$$

$$Upper\ quartile\ (Q3) = (N+1) * 0.75\,, \tag{3}$$

$$Interquartile\ range\ (IQR) = (Q3) - Q1\,, \tag{4}$$

Since our proposal method is based on a filter in multi-dimensional projections, in each column of data we will calculate the four statistical indicators Lower quartile (Q1), Middle quartile (Q2), Upper quartile (Q3) and interquartile range (IQR). The limits of the interval are defined as follows:

$$[Q1 - 1.5 * IQR\,; Q3 + 1.5 * IQR]\,, \tag{5}$$

Condition is necessary to declare extreme values as outliers in first time we filter data, that why we add

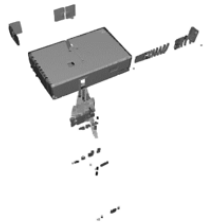
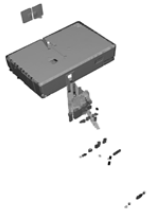
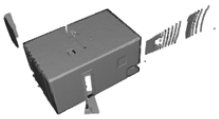
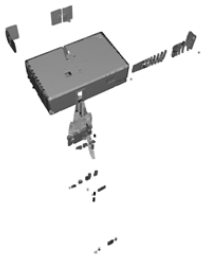
noisy building	X-axis filtration	Y-axis filtration	Z-axis filtration	combination of X, Y and Z axes
				
598603	2600	7249	0	9849

Figure 3. filtration according to different axis

The cleaning results for the dataset 'piece tours' shown in figure 3, show that the x-axis filtration was able to remove all the extreme points on the right and left of the part. On the other hand, when it comes to outliers at the back of the part, they were removed with y-axis filtration. The method finds it difficult to clean them along the z-axis, since in this case we do not have outliers that go beyond the ceiling or down to the floor.

4. EXPERIMENT AND RESULTS

4.1 Experiment setup

As previously stated, 3D scanning remains the most widely used method for acquiring the characteristic points of a previously constructed building. Indeed, this approach presents a certain number of advantages for the documentation of cultural heritage.[18] In particular, one can quote as advantages: a use as well for the interior scenes as for the external scenes, the speed of acquisition (2M points per second) or the precision of this one that can reach up to 1,9 mm per 10 m.[19] However, the quality of this digital model depends strongly on the specificity of the point cloud. Indeed, some characteristics must be taken into account if we want a quality 3D reconstruction. We can note in particular the density of points in the cloud, measurement errors and occlusions.

4.2 Result and description

In this part, a pre-processing step will be explored in order to clean the point clouds to lighten the database, by removing the outliers which represent a kind of noise in the point clouds that are the origin of a reflected laser on the mirrors and windows in particular and that influence the quality of the two following parts. The idea is to develop a new method based on static approaches. To test our method for cleaning outlier detection, we used 4 parts with different geometric shapes and types, buildings with rectangular or oval shape, modern or historical. The results of the cleaning of the parts illustrated in Table 1, show that the Statistical outlier and Removal outlier methods have well detected the outliers especially the scattered ones.

However, when it comes to outliers stuck to the main part, the method finds it difficult to clean them, since these points are considered as part of the part. However our methods was able to remove more outliers than the existing methods as it is mounted in the figure 4. The idea behind using this approach is to compute the box plot measure for each coordinate of the 3D point clouds (XYZ) in order to apply a filtering using an adaptive thresholds. The visual results are promising and the method cleans the buildings well and outperforms the Statistical removal and Radius Removal approaches as as shown in table 1.

Table 1. Number of data after the filtering process.

Dataset	Piece Tours	Export Brest	Export Labastide	Export Belves	
Nomber of outlier	8557	19347	9934	454298	Statistical
percentage of errors	1.42%	0.23%	0.010%	2.52%	removal
Nomber of outlier	7917	15514	8894	9047	Radius
percentage of errors	1.32 %	0.18%	0.096%	0.05%	removal
Nomber of outlier	9849	630350	7944	1155097	Our
percentage of errors	1.64%	7.69%	0.086%	6.42%	method
Total point	598603	8196905	9199416	17972687	

From the table in 1, it is clear that the two approaches 'Statistical Outliers Removal' and 'Radius Outliers Removal' from the Open3d library[20] for removing outliers in a point cloud. These two cleaning approaches, they worked quite well for simple buildings (modern rectangular rooms), however for historical buildings curved and non-rectilinear walls, they gave results that were not satisfactory.

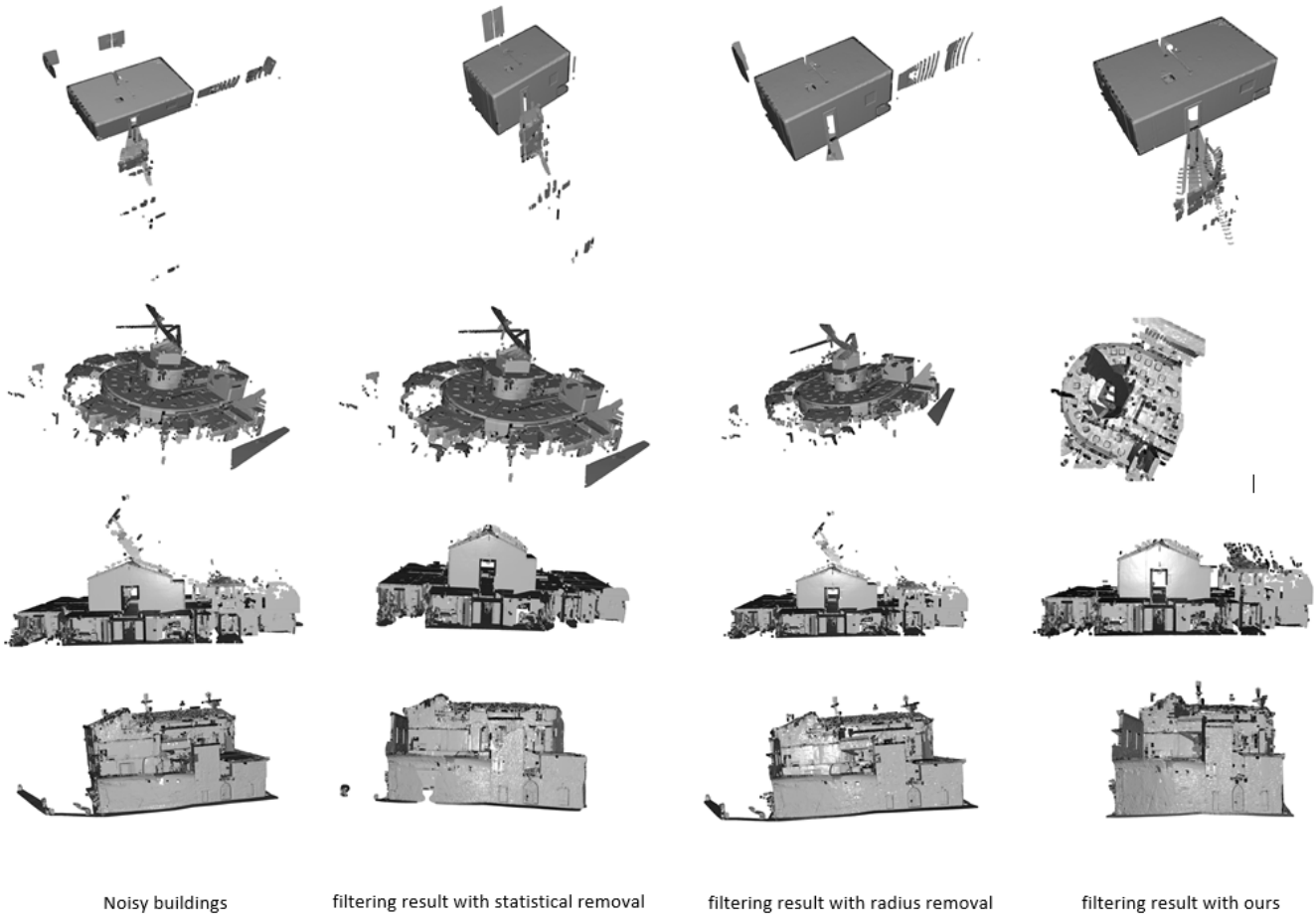


Figure 4. comparative study between the three approaches

5. CONCLUSION

In this paper, we present an efficient, simple and robust algorithm for eliminating outliers from often very noisy data. Our method reduces the number of errors and extraneous points in the input, which significantly improves the quality of the reconstruction while reducing computational and storage costs. The classical methods in the literature, "statistical removal" and "translation removal", often perform costly optimizations in terms of computation time and inefficiency of the results, which usually remove a significant amount of information from the scanned scene. this method is able to reduce these large and noisy point clouds as shown in comparison with the two existing methods in the experiments, which makes the reconstruction feasible and often even produces more accurate surface reconstructions that preserve many details. Therefore, we hope that our method opens the door to a new fundamental approach to 3D reconstruction. In this method we have presented a filtering method based on the box plot to remove outliers. This method is valid for any type of outlier, global outliers and collective outlier. Statistical methods are very effective on data sets with a known Gaussian distribution, while other methods are expensive in multivariate data and large diameters and require many parameters to be set. Even if several works are done on the detection of outliers each method still suffers from some drawbacks. New outlier detection methods would seem to influence the work reported in this article.

REFERENCES

[1] Hawkins, D. M., [*Identification of outliers*], vol. 11, Springer (1980).

[2] Hadi, A. S., Imon, A. R., and Werner, M., "Detection of outliers," *Wiley Interdisciplinary Reviews: Computational Statistics* **1**(1), 57–70 (2009).

[3] Dovoedo, Y. and Chakraborti, S., "Boxplot-based outlier detection for the location-scale family," *Communications in statistics-simulation and computation* **44**(6), 1492–1513 (2015).

[4] Han, X.-F., Jin, J. S., Wang, M.-J., Jiang, W., Gao, L., and Xiao, L., "A review of algorithms for filtering the 3d point cloud," *Signal Processing: Image Communication* **57**, 103–112 (2017).

[5] Chhabra, P., Scott, C., Kolaczyk, E. D., and Crovella, M., "Distributed spatial anomaly detection," in [*IEEE INFOCOM 2008-The 27th Conference on Computer Communications*], 1705–1713, IEEE (2008).

[6] Cai, Q., He, H., Man, H., and Qiu, J., "Iterativesomso: An iterative self-organizing map for spatial outlier detection," in [*International Symposium on Neural Networks*], 325–330, Springer (2010).

[7] Chawla, S. and Gionis, A., "k-means–: A unified approach to clustering and outlier detection," in [*Proceedings of the 2013 SIAM International Conference on Data Mining*], 189–197, SIAM (2013).

[8] Harris, P., Brunsdon, C., Charlton, M., Juggins, S., and Clarke, A., "Multivariate spatial outlier detection using robust geographically weighted methods," *Mathematical Geosciences* **46**(1), 1–31 (2014).

[9] Filzmoser, P., Ruiz-Gazen, A., and Thomas-Agnan, C., "Identification of local multivariate outliers," *Statistical Papers* **55**(1), 29–47 (2014).

[10] Chen, Y., Dang, X., Peng, H., and Bart, H. L., "Outlier detection with the kernelized spatial depth function," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **31**(2), 288–305 (2008).

[11] Papadimitriou, S., Kitagawa, H., Gibbons, P. B., and Faloutsos, C., "Loci: Fast outlier detection using the local correlation integral," in [*Proceedings 19th international conference on data engineering (Cat. No. 03CH37405)*], 315–326, IEEE (2003).

[12] Müller, E., Assent, I., Iglesias, P., Mülle, Y., and Böhm, K., "Outlier ranking via subspace analysis in multiple views of the data," in [*2012 IEEE 12th international conference on data mining*], 529–538, IEEE (2012).

[13] Gao, J., Liang, F., Fan, W., Wang, C., Sun, Y., and Han, J., "On community outliers and their efficient detection in information networks," in [*Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*], 813–822 (2010).

[14] Kou, Y., *Abnormal pattern recognition in spatial data*, PhD thesis, Virginia Tech (2006).

[15] Griffin, T. W., Dobbins, C. L., Vyn, T. J., Florax, R. J., and Lowenberg-DeBoer, J. M., "Spatial analysis of yield monitor data: Case studies of on-farm trials and farm management decision making," *Precision Agriculture* **9**(5), 269–283 (2008).

[16] Filzmoser, P., Maronna, R., and Werner, M., "Outlier identification in high dimensions," *Computational statistics & data analysis* **52**(3), 1694–1711 (2008).

[17] Mendez, D. and Labrador, M. A., "On sensor data verification for participatory sensing systems," *Journal of Networks* **8**(3), 576 (2013).

[18] Galeazzi, F., "3d recording, documentation and management of cultural heritage," (2017).

[19] Azhar, S., Khalfan, M., and Maqsood, T., "Building information modeling (bim): now and beyond," *Australasian Journal of Construction Economics and Building, The* **12**(4), 15–28 (2012).

[20] Verma, R. and Verma, A. K., "A clustering and outlier detection scheme for robust parametric model estimation for plane fitting," in [*Applied mechanics and materials*], **789**, 770–775, Trans Tech Publ (2015).