General Clustering Survey

Literature Review

Jemma Sundelson

University of Cape Town

Department of Computer Science

Cape Town South Africa

sndjem001@myuct.ac.za

ABSTRACT

Clustering analysis is a complex, yet interesting problem. It is a tool that is used in many aspects of research as well as in a variety of fields to analyze and extract data. Specifically, so the full potential from gathered information is gained. In this literature review, existing clustering techniques are examined to find the most suitable method to group points in a point cloud based on a similarity measure to generate meaningful results. Therefore this paper surveys clustering as a whole as they pertain to point cloud clustering as well as assesses and contrasts the main clustering methods such as partitional, hierarchical, and, density-based clustering. The papers examined in this literature review show that there is no universal technique that performs well in all contexts. However, there are techniques that can effectively classify point clouds leading to satisfactory results such as K-means clustering, Fuzzy c-means clustering, Hierarchical clustering as well as DBSCAN.

KEYWORDS

Point-clouds, Clustering, Machine Learning

1 INTRODUCTION

As the world advances technologically, many complex problems arise and need to be solved. Vast amounts of data are being collected each day therefore the process of data mining has become of great importance. Cluster analysis is a widely used tool to analyze and dissect these large data sets to obtain meaningful, new information. A cluster is a set of similar, unlabeled patterns. Data clustering is an unsupervised learning technique that identifies natural groupings within data based on some similarity measure [33]. i.e., unlabeled samples are assigned to groups [25]. Data clustering can be performed in a variety of different contexts and will be explained thoroughly in sections to follow. Given the power of clustering, it has been applied to complex domains such as the segmentation (clustering) of cultural heritage point clouds.

The preservation of Cultural Heritage sites such as art, archaeology, and architecture are of great importance to society as it is a means of transmitting cultural heritage to future generations [25, 36]. An important way to achieve sustainability of these cultural heritage sites is through digital technologies [36]. Many new techniques for preservation are being explored, for example, the Zamani project at the University of Cape Town uses terrestrial laser scanning (TLS) to record point cloud data for several cultural heritage sites within Africa and other continents [25]. A point cloud is a set of points defined by a coordinate system [25]. 3D point clouds have an X, Y, and Z coordinate therefore they can represent the shape, size, position, and orientation of objects in space [23]. Figure 1 below depicts the laser scanning pipeline which is the process of creating a 3D computer model from laser scan data [28].

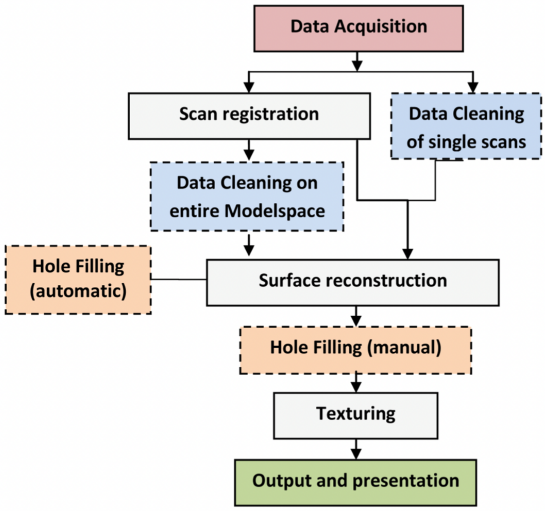

Figure 1: The laser scanning pipeline [28]

In the Data Acquisition stage of the laser scanning pipeline, the recordings from the scanning are stored as point clouds which are used at a later stage to generate 3D models [25]. Secondly, during scan registration, all scans are transformed into a single uniform coordinate system [28]. The last stage that will be discussed is the cleaning of point clouds. This process involves removing 'noise'. Noise includes all irrelevant features such as vegetation, animals, people, and so on. This task is often done manually and can be performed on range images before registration or on point clouds after registration [20, 27]. However, due to very large amounts of data, it is time-consuming and labor-intensive. For large heritage sites, it can take an experienced individual between 30 and 120 minutes [20]. Thus, to address this problem, machine learning techniques such as clustering are explored to semi-automate this process.

This paper will survey unsupervised learning used to recognize patterns in unlabeled data, more specifically different clustering techniques. This is relevant as clustering methods group nearby points with similar features by forming segments based on grouping extracted point features [25]. Therefore, if we can cluster points in 3D scans accurately, we can interpret the 3D point set. This can then be cleaned by removing 'noise' clusters. Once this is done, clean 3D output models can be generated which will contain only desired geometric data. Due to the research done in this area and the extensive, comprehensive different clustering techniques, this literature review will only survey the most commonly used clustering techniques such as K-means, Fuzzy c-means, Hierarchical clustering, and, DBSCAN.

2 CLUSTERING

According to H.P. Kriegal et al [16], in the last decade, new techniques have been found to cope with new challenges that have arisen from the modern capabilities of automatic data generation. Extensive data is being produced and needs to be examined by data mining methods to gain the full potential from gathered information. There are different data mining methods namely descriptive statistics, visual data mining methods as well as machine learning methods [5]. This literature review aims its focus on a machine learning technique that uses unsupervised learning, namely clustering. Clustering is an important technique used to classify data as it organizes large amounts of information into meaningful clusters [38]. This is done by grouping elements (images, point clouds, or meshes) that are similar under some metric. Patterns in one cluster are more similar to one another than they are to patterns in another cluster[15]. This is shown below in Figure 2.

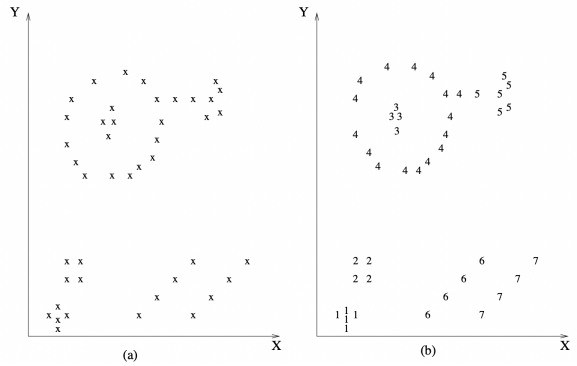

Figure 2: Data clustering [15]

Figure 3 represents a broad overview of the different stages of the clustering process [15].

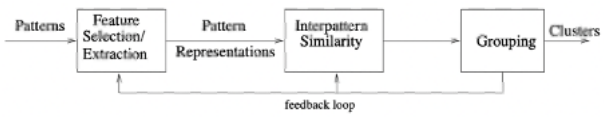

Figure 3: The clustering process [15]

Patterns represent the number, type, and scale available to the clustering algorithm [15]. Feature selection refers to the selection of the most relevant features which will be used in the clustering process. Elegant feature selection can largely decrease the workload of the clustering process [37]. Inter pattern similarity represents the relationship between data and lastly, the grouping stage depicts the different clustering techniques, i.e., the way the clusters are formed which can be done in multiple different ways.

Moreover, clustering is an important technique used to classify or detect outliers, simplify data for further analysis or, to visualize data. Therefore, it is clear that clustering has many purposes and can be used in various fields such as engineering, medicine, marketing, and archaeology. For example, clustering is important for biologists who want to find genes with similar properties, in medical research to find genetic relationships and disorders, detecting metabolic diseases at the earliest possible moment for newborns and for customer recommendation systems [16]. Many different clustering techniques have been discovered to perform tasks as above, the main ones being partitional, hierarchical, and density-based clustering which will be discussed in-depth in the following sections.

However, all research on specialized approaches to clustering data is relatively new [16]. it is still an unanswered question as to which clustering technique is best to use. The performance of these methods tends to depend heavily on what the problem trying to be solved is as well as the quality of the training data [25]. For example, the k-means algorithm is attractive to use for large data sets as it is computationally inexpensive but if a user tried to use hierarchical clustering on that same set of data it may fail as hierarchical clustering is computationally expensive. Some methods tend to perform better than others leading to satisfactory results, yet these still come with drawbacks, for example, the user needing to specify input parameters which can be a difficult task to solve. Additionally, another difficulty to overcome when using various clustering techniques is identifying noise (irrelevant features) that heavily influences the clustering result. Therefore, one of the main challenges involved in clustering is determining which features are relevant and which are not so they can be discarded, and meaningful results can be obtained.

Thus, Clustering is an interesting, complex problem. It is a powerful technique that yields important and new information. The following sections will describe the main clustering techniques.

2.1 Partitional Clustering Techniques

Partitioning algorithms work by defining an initial number of groups and then iteratively reallocating objects among them [18] until they reach a point of convergence. This technique determines all clusters at the same time [33].

2.1.1 K-means Algorithm. The K-means algorithm is the most widely used partitional clustering technique [11]. The user specifies the number of clusters (K), that they want, ahead of time and the algorithm works to calculate the means of data points. The algorithm groups nearby points with similar features by assigning each

point to the cluster whose centroid (the average of all points in a cluster) is nearest [18].

Salah et al. [30] provides the pseudo code of the k-means algorithm:

(1) Randomly allocate each data-point to one of the k clusters and initialize K centroids for each cluster

(2) allocate each data point to the closest centroid

(3) Re-compute the centroids using the current cluster memberships

(4) If a convergence criterion is not met – go to step 2

(5) Steps 2-4 are repeated until a convergence criterion is met.

Assigning data points to the closest centroid requires that a similarity (distance) metric be defined. The convergence criterion is when the positions of the cluster centroids are no longer changing position (in the data coordinate space)

The user must specify K beforehand therefore a commonly used method for finding the optimal K value is the Elbow Method [13]. This method works by running the k-means algorithm on the data for a range of values for K. The within-cluster sum of square is calculated for each k-value and decreases as the number of clusters increases. In Figure 3, the graph forms an elbow shape, showing the optimal K value is at this point where the variation within clusters is small and stable.

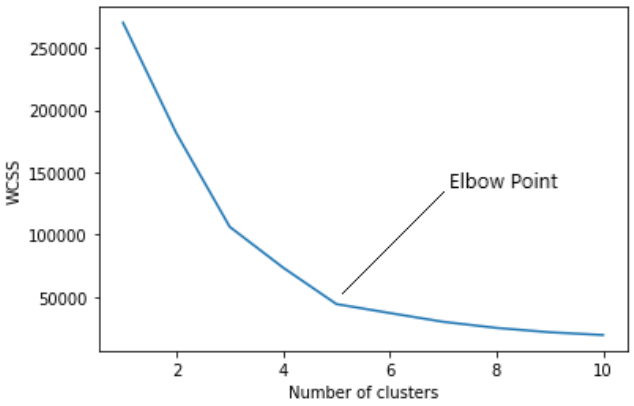

Figure 4: Elbow method, clusters for data = 4 [13]

2.1.2 Fuzzy c-means Algorithm. The Fuzzy c-means algorithm is a version of the K-means algorithm above. It was developed by Dunn, 1974 [8]. and improved by Bezdek, 1981 [24]. The goal of the Fuzzy c-means Algorithm is to efficiently model ambiguous unsupervised patterns [6]. A key aspect of this algorithm is that each data point can be assigned to more than one cluster. Just like the K-means algorithm, the number of clusters must be specified by the user before starting the computation.

2.2 Hierarchical Clustering

The hierarchical clustering technique is used when there is an implicit hierarchical structure within the data. These algorithms generate a cluster tree (dendrogram) by using heuristic splitting or merging techniques [14, 33].

A dendrogram is a tree in which categories and subcategories are grouped at different levels. According to Madhulatha T.S. et al [18]. there are two approaches to building a dendrogram. The first is Agglomerative which is a bottom-up technique that uses merging to build cluster trees. The second is Divisive which is a top-down technique that uses splitting to generate clusters. Agglomerative hierarchical algorithms begin by allocating each pattern to one cluster, and then the two most similar clusters are merged. This step is repeated until all patterns are assigned to a single cluster [22, 35]. In contrast, divisive hierarchical algorithms start with all patterns allocated to a single cluster, then each cluster is split in each stage until each cluster has one pattern [22].

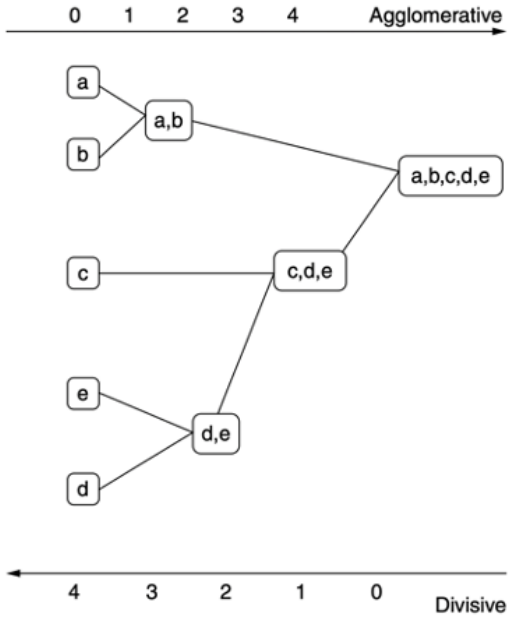

Figure 5: Examples of a Hierarchical tree structure [9]

2.3 Density-Based Clustering

Density-based clustering can detect and manage arbitrary-shape clusters as well as noise as it works by forming clusters based on the dense areas that are separated by skimpy areas [4].

2.3.1 DBSCAN. DBSCAN (Density-Based Spatial Clustering of Applications with Noise) [29]. algorithm is based on the denseness of clusters. Regions with a high density of points describe the existence of clusters whereas regions with a low density of points indicate clusters of noise or outliers[18]. The density of any point is determined by the number of points in its immediate vicinity, known as the point's neighborhood [4]. The dense neighborhood is defined by two user-specified parameters: the radius (e) of the neighborhood (e-neighborhood), and the number of the items in the neighborhood (MinPts) [4].

The DBSCAN process is performed as follows:

(1) Choose a random point p

(2) investigate if the e-neighborhood of the point contains at least MinPts points[3].

(3) If p is a core point (a point with a cardinality of at least MinPts [4]) then a cluster is formed

(4) If p is not a core point it is considered a noise point [4] and DBSCAN iteratively moves to the next point in the database [17].

(5) Repeat the process until all points are explored and no new point can be added to any cluster [3]

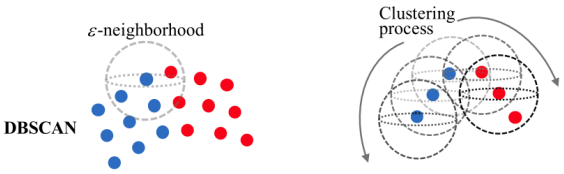

Figure 6: DBSCAN e-neighborhood and clustering process [7]

3 CONTRASTING CLUSTERING METHODS ON POINT CLOUDS

3.1 Point Clouds

Point clouds are processed to generate high-resolution 3D models. For example, LiDAR (Light Detection And Ranging) scanning is used as it can capture points over a long distance by the scanner probing the subject with a laser light and times the round-trip of each pulse of light [25]. Two different types of scanning's use LiDAR namely airborne laser scanning (ALS) and terrestrial laser scanning (TLS). ALS and TLS data are conveyed as a very large dataset of point clouds [26]. and various methods are used to extract features such as a crack on a wall, slope of a surface, and so on. Although LiDAR scanning is very useful, it also comes with drawbacks namely lower accuracy, scattered point effect due to the averaging of two different locations of one pulse as well as decreased density of a point as distance increases [25]. Additionally in more rural areas, often these LiDAR scanning's will pick up unwanted features such as vegetation and animals, therefore, misrepresenting the generated 3D model.

3.2 Point cloud features

To obtain meaningful 3D representations of objects, we need to extract features from the point clouds. Features can be described as a piece of information used to describe an object or part of an object [1]. The extraction of features allows data to be separated into clusters that share similar properties (segmentation) as well as label each segment with a class to give some semantic to the segment (classification) [32].

3.3 Point cloud segmentation

Point cloud segmentation is the process of classifying point clouds into different regions. The points in the same region will have the same properties [21]. However, this process is not a simple task as point cloud data contains irrelevant features (noise), they are disorganized as well as lack structure. Additionally, due to different angles of the scanner, the sampling density of points can also be uneven [21].

There are many different methods for 3D point cloud segmentation such as Edge-based methods, Region-based methods, model-based methods, Graph-based methods [12]. and clustering. However only clustering methods will be explored in-depth in the following section.

3.4 Experiments

Grouping similar points are efficiently done by clustering techniques. This section surveys different clustering techniques on point clouds by reviewing different experiments.

Shan et al. [31]. utilized clustering specifically to try to find which points belong to a specific roof segment. LiDAR point clouds were collected from the roof of a building. Noise points were eliminated early in the clustering process so that their representation was as accurate as possible. The clustering technique used was the K-means algorithm because it is simple and most commonly used. Additionally, to select the most accurate K, the elbow method was implemented.

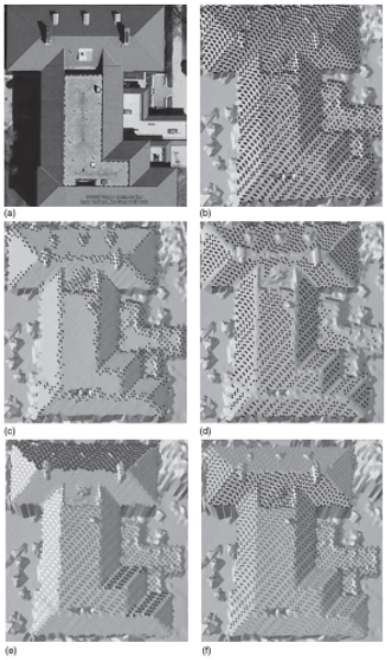

Figure 7: roof LiDAR points clustering and segmentation. (a) image from Google, (b) LiDAR points, (c) break line points, (d) planar points, (e) clustered points and (f) planar segments. [31]

Figure 5 shows the results of the experiment. The researchers found that almost all break line points are depicted correctly, and the performance is satisfactory. Additionally, small roof attachments were also well detected as break line points. However, separation is not always perfect as there exists over-removal of planar points. But this shouldn't be a concern due to the large number of available LiDAR points on the roof.In addition, the authors found the results of employing the k-means method were adequate and show promise for modest and complicated constructions. Even though there were outliers like trees and break line locations, the algorithm worked effectively. These findings can be related to the segmentation of point clouds on cultural heritage sites. Cultural Heritage site point clouds are also rough, noisy (contain outliers), and abundant. Therefore as the k-means algorithm performed successfully in this context, there is potential for it to yield satisfactory results in the context of segmentation of cultural heritage sites.

One thing to keep in mind is that removing outliers as soon as possible is critical because they might have a significant impact on the k-means experiment. Furthermore, one of the drawbacks of this method is that the user must define k before the experiment can begin. In this situation, k stood for the number of directional roof planes in a building, which is generally unknown prior to clustering

Additionally, Raffaele Albano investigates roof segmentation for 3D reconstruction from aerial LiDAR point clouds [2]. The data set was captured over downtown Toronto (Canada). Optech's ALTM-ORIONM was used to acquire the ALS data. When constructing 3D building models, the segmentation process aims to find out which LiDAR point clouds belong to which particular roof segment. The fuzzy c-means algorithm was tested on different types of urban development. In this method, data points do not belong to only one cluster.

Segmentation Approach	Average RMSE (m)	Average Dist. Mean (m)	Average Dist St. Dev. (m)	Average Computational Time (sec)
Clustering	3.81	3.20	2.08	17

Figure 8: : Average performances of the geometry accuracy on 58 buildings over downtown Toronto (Canada). [2]

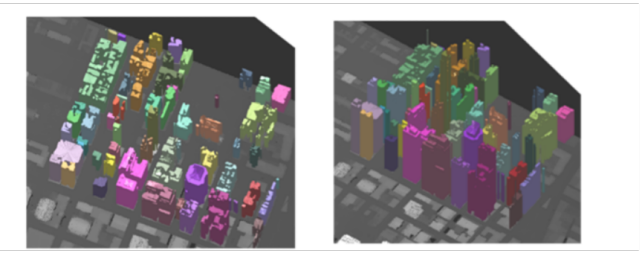

Figure 9: : Results of 3d buildings reconstruction over downtown Toronto (Canada) using the clustering roof segmentation approach [2]

The findings reveal that the fuzzy c-means algorithm achieves acceptable performance metrics in terms of geometry correctness, indicating that this algorithm's adaptability allows it to be employed in a variety of situations. Therefore its adaptability may make it suitable for the segmentation of cultural heritage sites. However, nothing was mentioned about outliers and hence we do not know how well the algorithm would perform if outliers were considered.

Moreover, the extraction and repairment of gaps in point clouds of power lines, using the hierarchical clustering technique is investigated by Fan, Yongzhao, et al. [10].

Power lines are extracted using a hierarchical clustering approach. Gaps will emerge in scans as a result of variable point density, influencing power line detection through over clustering and insufficient extraction.

The investigation was carried out using terrestrial laser scanning point cloud data and was measured near the Land and Resources College in Hannan District of Wuhan, China. It was found that Gaps are corrected based on neighborhood relations of power line candidates during hierarchical clustering, and fixed gaps generate continuous neighborhood relations that allow the clustering procedure to be executed.

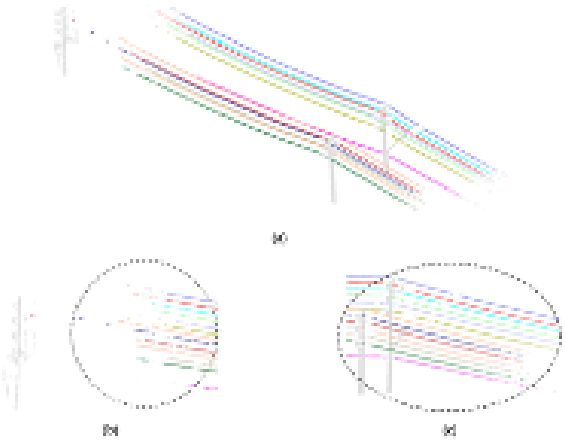

Figure 10: : The results of power line extraction. (a) Each single power line is successfully extracted from preprocessed data directly and shown in different colors. (b,c) Clustering with the gaps. [10]

The power lines are recovered from pre-processed data and separated into individual lines in Figure 7, suggesting that the power lines are grouped correctly. The experiment's findings show that the hierarchical clustering method efficiently solves the problem of gaps and that individual power lines may be recovered straight from pre-processed data without any prior condition. This is an important finding as gaps can also appear in cultural heritage point cloud data due to incorrect TLS as well as obstructions.

Another experiment to be examined is the rockfall detection from terrestrial lidar point clouds using DBSCAN clustering technique conducted by Tonini et al [34].

The study site was a cliff in Puigceros, Spain. This cliff features numerous rockfalls each year, making it ideal for the research study.

The researchers began by identifying the various cluster events (rockfalls) and labeling each feature according to the classification results. DBSCAN, a density-based method, was used. DBSCAN was chosen because it can find clusters of any shape in a huge spatial database with only a few parameters. The minimal number of points (MinPts) inside a maximum distance (eps) surrounding each randomly chosen point in the data set are the input parameters required. If each point p meets the MinPts requirement, the algorithm explores it and labels the points at eps distance. The pre-filtered data set made it possible to classify feature points as belonging to different rockfall occurrences. The feature points allocate well-shaped rockfall events, however, they are not assigned to a particular rockfall event implying, they are not individually labeled. As a result, DBSCAN was utilized to fix this issue. The experiment's results are displayed below.

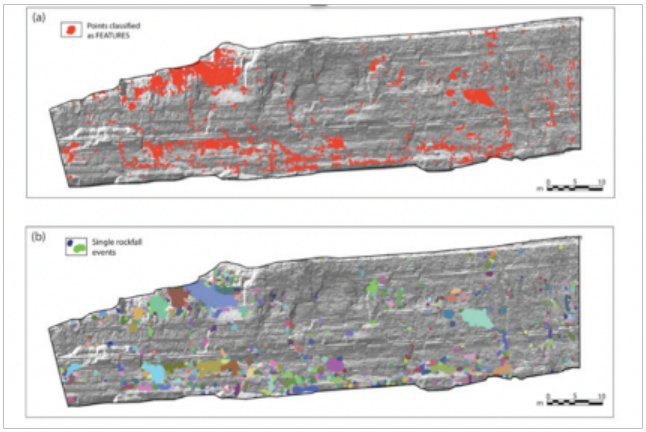

Figure 11: Colored dots represent single rockfall events resulting from DBSCAN [34]

Figure 8 depicts that the DBSCAN method allowed labeling the denser aggregations of feature points as belonging to a single rockfall event and removing the residual noise features.

The DBSCAN technique, according to the research article, was particularly effective for categorizing each feature point as belonging to a single rockfall event, allowing single rockfalls to be mapped and extracted. Therefore, since the DBSCAN technique performed well in this context, it could yield promising results in identifying clusters of different shapes and grouping geometric features in the context of cultural heritage sites. However, every algorithm has limitations, and one of DBSCAN's is that it significantly relies on the input parameters MinPts, and eps cannot be entirely unsupervised because the user must examine the results and sometimes alter these parameters as a result.

4 DISCUSSION

The above experiments explored different clustering techniques and showed that there are pros and cons to each method. All four algorithms used above achieved successful results yet there were limitations to each. For example, the k-means algorithm has the advantages of being simple to implement and having a time complexity of O(Np), making it attractive to use for big data sets [33]. It is also order-independent, which means that for a given initial seed set of cluster centers, it generates the same partition of data regardless of the order in which the patterns are given to the algorithm [19]. However, one disadvantage is that the user must first determine K (the number of clusters) before utilizing the algorithm, which is not always an easy to determine.

Moreover, the fuzzy c-means algorithm has some advantages over the k-means algorithm in that each cluster is assigned a pattern with some degree of membership making it more suitable for real applications where there are some overlaps between clusters in the data [22]. However, just like in k-means, the user must specify the number of clusters beforehand.

Furthermore, DBSCAN performs well in identifying clusters of any shape, and additionally, the user does not need to define the number of clusters beforehand. However, there are also drawbacks to using DBSCAN. DBSCAN does not have flexibility in defining epsilon or minimum points so it can never take a step larger than epsilon (the local radius) [17]. And therefore, the success of the clustering process is heavily dependent on how the user defines the input parameters above.

Lastly, Hierarchical clustering is advantageous in the sense that the number of clusters does not need to be specified beforehand and they are independent of initial conditions [22]. Additionally, Hierarchical Algorithms are more versatile [1]. than other clustering techniques. However, there are also limitations to the algorithm. The first being, that it is very computationally expensive [33]. therefore, it is not fitting with large data sets. The second being, that the algorithm is static. Once a merge/split is done, it cannot be undone [3]. meaning, it is not possible to correct possible previous miscalculation. [37]

These findings can now be applied to the complex domain of segmentation of cultural heritage point clouds. Point cloud data from cultural heritage sites tend to be noisy and rough. However, if 'noise' points can be removed early in the clustering process and the number of clusters, is easy to determine, then k-means clustering will be useful as it is easiest to implement. In addition, since there is abundant points in point cloud data for cultural heritage sites, the k-means algorithm will be attractive due to its low computational complexity. Fuzzy c-means algorithm should also be considered in this respect as it has similarities to the k-means algorithm however, also allows for the overlapping of clusters. Therefore making it more flexible. Additionally DBSCAN performs successfully in classifying points within a point cloud while discarding noise, therefore could yield promising results in the context of cultural heritage sites. Nevertheless, requires a large amount of memory to perform precisely. Moreover, results showed that the hierarchical clustering technique performed well in detecting gaps in powerlines. Point cloud data from cultural heritage sites may contain gaps if the laser scanning is not performed correctly, thus, a hierarchical clustering method has the potential to perform well in this context. However the user must be aware of the large computational complexity of this method. If the point cloud data is too large, the user may run into problems.

In conclusion, it is clear that the user must choose a clustering technique that is relevant to the context of their data. It is important to assess the advantages and disadvantages of using each algorithm when making an educated decision of which clustering technique to choose so successful results can be obtained. The findings above show successful applications of cluster analysis however there are still many unanswered questions in this domain due to the existence of uncertain factors [37]. We can conclude there is no universal technique that can be applied to all contexts of data.

5 CONCLUSION

Existing research shows that clustering is a powerful, yet challenging tool used to solve many complex problems, and when used correctly, successful, important results can be obtained. Moreover, Clustering techniques can be performed in a wide variety of contexts. There are advantages to using each clustering method however every technique has its own limitations. Therefore, it is important that a user has a deep knowledge of the different methods as well as assesses the trade-offs of each technique to make an informed, educated decision for their particular data set to obtain satisfactory results. This literature review has assessed different clustering techniques and can conclude that clustering applications are suitable to use in the domain of cultural heritage sites to effectively cluster point cloud data. However, it is clear that data clustering is a subjective process which often makes it difficult to perform as the performance of these clustering techniques depends on the quality of the training data and the context of the problem trying to be solved. Therefore, through contrasting and reviewing several clustering techniques, we can conclude that there is no universal clustering technique that can be applied to all contexts of data. A technique that works well in one aspect may fail with other data sets.

REFERENCES

[1] Abbas, O. A. Comparisons between data clustering algorithms. *International Arab Journal of Information Technology (IAJIT) 5*, 3 (2008).

[2] Albano, R. Investigation on roof segmentation for 3d building reconstruction from aerial lidar point clouds. *Applied Sciences 9*, 21 (2019), 4674.

[3] Amini, A. *An adaptive density-based method for clustering evolving data streams.* PhD thesis, University of Malaya, 2014.

[4] Amini, A., Wah, T. Y., and Saboohi, H. On density-based data streams clustering algorithms: A survey. *Journal of Computer Science and Technology 29*, 1 (2014), 116–141.

[5] Bandaru, S., Ng, A. H., and Deb, K. Data mining methods for knowledge discovery in multi-objective optimization: Part b - new developments and applications. *Expert Systems with Applications 70* (2017), 119–138.

[6] Baraldi, A., and Blonda, P. A survey of fuzzy clustering algorithms for pattern recognition. i. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 29*, 6 (1999), 778–785.

[7] Chen, H., Liang, M., Liu, W., Wang, W., and Liu, P. X. An approach to boundary detection for 3d point clouds based on dbscan clustering. *Pattern Recognition 124* (2022), 108431.

[8] Dunn, J. C. A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters.

[9] Everitt, B., Landau, S., Leese, M., and Stahl, D. Wiley series in probability and statistics, 2011.

[10] Fan, Y., Zou, R., Fan, X., Dong, R., and Xie, M. A hierarchical clustering method to repair gaps in point clouds of powerline corridor for powerline extraction. *Remote Sensing 13*, 8 (2021), 1502.

[11] Goyal, A., Tirumalasetty, S., Hossain, G., Challoo, R., Arya, M., Agrawal, R., and Agrawal, D. Development of a stand-alone independent graphical user interface for neurological disease prediction with automated extraction and segmentation of gray and white matter in brain mri images. *Journal of Healthcare Engineering 2019* (2019).

[12] Guimarães, N., Pádua, L., Marques, P., Silva, N., Peres, E., and Sousa, J. J. Forestry remote sensing from unmanned aerial vehicles: A review focusing on the data, processing and potentialities. *Remote Sensing 12*, 6 (2020), 1046.

[13] Gujski, L., di Filippo, A., and Limongiello, M. Machine learning clustering for point clouds optimisation via feature analysis in cultural heritage. *International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences* (2022).

[14] Hamerly, G. J. *Learning structure and concepts in data through data clustering.* University of California, San Diego, 2003.

[15] Jain, A. K., Murty, M. N., and Flynn, P. J. Data clustering: a review. *ACM computing surveys (CSUR) 31*, 3 (1999), 264–323.

[16] Kriegel, H.-P., and Kr, P. oger, and a. zimek,"clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering,". *ACM Trans. Knowl. Discovery Data 3*, 1 (2009), 1–58.

[17] Kurumalla, S., and Rao, P. S. K-nearest neighbor based dbscan clustering algorithm for image segmentation. *Journal of Theoretical and Applied Information Technology 92*, 2 (2016), 395.

[18] Madhulatha, T. S. An overview on clustering methods. *arXiv preprint arXiv:1205.1117* (2012).

[19] Maimon, O., and Rokach, L. Data mining and knowledge discovery handbook.

[20] Mulder, R. Accelerating point cloud cleaning. Master's thesis, University of Cape Town, 2017.

[21] Nguyen, A., and Le, B. 3d point cloud segmentation: A survey. In *2013 6th IEEE conference on robotics, automation and mechatronics (RAM)* (2013), IEEE, pp. 225–230.

[22] Omran, M. G., Engelbrecht, A. P., and Salman, A. An overview of clustering methods. *Intelligent Data Analysis 11*, 6 (2007), 583–605.

[23] Oniga, V.-E., Breaban, A.-I., Pfeifer, N., and Diac, M. 3d modeling of urban area based on oblique uas images—an end-to-end pipeline. *Remote Sensing 14*, 2 (2022), 422.

[24] Peizhuang, W. Pattern recognition with fuzzy objective function algorithms (james c. bezdek). *Siam Review 25*, 3 (1983), 442.

[25] Pocock, C. 3d scan campaign classification with representative training scan selection. Master's thesis, Faculty of Science, 2019.

[26] Razak, K. A., Straatsma, M., Van Westen, C., Malet, J.-P., and De Jong, S. Airborne laser scanning of forested landslides characterization: Terrain model quality and visualization. *Geomorphology 126*, 1-2 (2011), 186–200.

[27] Rüther, H., Held, C., Bhurtha, R., Schröder, R., and Wessels, S. Challenges in heritage documentation with terrestrial laser scanning. In *Proceedings of the 1st AfricaGEO Conference, Capetown, South Africa* (2011), vol. 30.

[28] Rüther, H., Held, C., Bhurtha, R., Schroeder, R., and Wessels, S. From point cloud to textured model, the zamani laser scanning pipeline in heritage documentation. *South African Journal of Geomatics 1*, 1 (2012), 44–59.

[29] Saji, B. In-depth intuition of k-means clustering algorithm in machine learning. *Analytics Vidhya: https://www. analyticsvidhya. com/blog/2021/01/in-depth-intuition-of-k-meansclustering-algorithm-in-machine-learning/,(Accessed on 6/11* (2021).

[30] Salih, H. A., Ahmed, A. S., and Jameel, J. Q. Development of a decision support system for urban planning by using k-means++ algorithm. *Al-Mustansiriyah Journal of Science 31*, 3 (2020), 78–88.

[31] Shan, J., and Sampath, A. Building extraction from lidar point clouds based on clustering techniques. *Topographic laser ranging and scanning: principles and processing* (2008), 423–446.

[32] Shao, J., Zhang, W., Shen, A., Mellado, N., Cai, S., Luo, L., Wang, N., Yan, G., and Zhou, G. Seed point set-based building roof extraction from airborne lidar point clouds using a top-down strategy. *Automation in Construction 126* (2021), 103660.

[33] Swarndeep Saket, J., and Pandya, S. An overview of partitioning algorithms in clustering techniques. *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) 5*, 6 (2016), 1943–1946.

[34] Tonini, M., and Abellan, A. Rockfall detection from terrestrial lidar point clouds: A clustering approach using r. *Journal of Spatial Information Science*, 8 (2014), 95–110.

[35] Turi, R. Clustering-based colour image segmentation (ph. d. thesis). *Monash University, Melbourne* (2001).

[36] Walston, S. The preservation and conservation of aboriginal and pacific cultural material in australian museums. *ICCM bulletin 4*, 4 (1978), 9–21.

[37] Xu, R., and Wunsch, D. Survey of clustering algorithms. *IEEE Transactions on neural networks 16*, 3 (2005), 645–678.

[38] Zhao, Y., and Karypis, G. Evaluation of hierarchical clustering algorithms for document datasets. In *Proceedings of the eleventh international conference on Information and knowledge management* (2002), pp. 515–524.