Point Cloud Data Organization Algorithm Based on Clustering

Kun Zhang¹,², Weihong Bi¹, Xiaoming Zhang², Xinghu Fu¹, Kunpeng Zhou¹, Li Zhu¹

¹College of Information Science & Engineering, Yanshan University, Qinhuangdao, Hebei, China
²College of Information Science & Engineering, Hebei University of Science & Technology, Shijiazhuang Hebei, China
euphkun@163.com, bwhong@ysu.edu.cn, zhangxiaom@hebust.edu.cn, fuxinghu@ysu.edu.cn, kunpeng032@126.com, gaudothsin@sina.com

Abstract. With the development of computer hardware, the 3D scanner is cheaper, and more and more 3D data can be got from 3D scanner. The data can access as the point cloud data. So, the processing for large-scale point data has been as the new branch of computer graphics. This paper provided a new algorithm for point cloud organization based on improved Kmeans. The point cloud can be store as a tree. At last, the algorithm is verified by the experiment results.

Keywords: point cloud, Kmeans clustering algorithm, data density

1 Introduction

With the development of 3D scanner, the point data can be got conveniently. Though CAD(computer aided design), point data[1] can be as an important data form. Rather, it would be as an effective and popular data form in reverse reconstruction. In general, the reverse reconstruction based on point cloud data requires more than four steps. Firstly, the data can be got from 3D scanner, and we have to eliminate the data noise, reduce the data scale, and complete surface reconstruction. What more, besides the point data processing, the data organization is an important thing. The convenient data access can save the query time on the data adjacent domain, and faster human-interactive and dynamical update.

We take the methods for point data organization algorithm as two main categories. 1) Manage data based on sub-space. The point cloud data will be organized by the sub-space, we can call it as space division [2]. For example, grid method[3], quadtree and octree subdivision method[4], KD-tree[5]. 2) Space database in existed software. These software, for example Arcgis, Terrosolid, Polyworks and RealWorks[6], will place the emphasis on data processing, so during the processing ,the point data would be seen as one single file, or the data from one device as one single file. So the software for reverse reconstruction would spend most of the time to process data, such as find neighbor data, data domain.
2 Related Work

The Kmeans algorithm based on density constraints is suitable for a large-scale and uneven distributed point data. Because then part of the point data, which we concerned, is a small set of whole, with concentrated distribution. And we can get the hit from this feature. That is if the distance of points as effecting factors involved into data organization, the reverse reconstruction algorithm speed would be speed up. Hence, during the point cloud processing, we join the clustering algorithm into it.

Firstly, the scatter points can be divided into different sets as which distance of each other, and the data density as the end of clustering algorithm. Then, the center of the cluster sets and the set density value can be stored as cluster identification data. At Last, the cluster sets can be organized as a cluster tree.

For data domain access conveniently and speed up human-computer interaction process, the Kmeans clustering algorithm by the data itself feature has been took into account in this paper. Besides, we have provided a new point data organization algorithm based on improved Kmeans clustering.

In order to suit for the large scale point cloud and get the relevance of point data itself, we improved traditional Kmeans algorithm. Then, Kmeans algorithm with density constraints will produces a new form of result. What's more, one point would not be convergent to one cluster. So, there will be different results about cluster set, we call these as the intersection cluster, adjacent cluster and isolation cluster.

3 Data Organization Algorithm Based On Improved Kmeans

In order to find neighbor data conveniently, this paper set the end condition of data density, based on Kmeans clustering. And this makes some data cannot be reach convergence on one cluster. Lucky, the data set which convergence on two or more cluster, means these cluster set is adjacent set and the data is adjacent data.

As mentioned above, the improved algorithm pseudo-code is as follows:

Step1. The center of the clusters randomly set $\{\mu_k | k = 1...K\}$;

Step2. While $(P_k > \sigma) \lor (\text{Num Iteration} > \text{Max})$)

Step3. Create cluster sets based on $\{\mu_k | k = 1...K\}$;

Step4. Compute the center of the cluster again, and update $\mu_k$;

Step5. Compute density of data again, and update $P_k$;

As it described above, there will be different results set of cluster, which can be divided into three parts. Among that, the subset of intersection clusters and adjacent clusters are important for neighbor of data domain. The result about data correlation can be stored as a forest, simply. However, as we all know, this method is difficult for building organization form and searching.
The real threshold value of density $\sigma$ can be computer as the parameter of the 3D scanner and the movable platform, and we call it as uncertainty data, which has been discussed in the reference [7].

The contour data carry out by sort algorithm, the processes can be described as the expression:

$$\text{sort}(\min(c_{i,d}), \min(c_{i,d}), \max(c_{i,d}), \max(c_{i,d}))$$

Based on the result of sort algorithm, the adjacent and isolate relation of the clusters could be find out. Among them, the isolate clusters can be organized as an isolate table.

The data organization algorithm pseudo-code of cluster tree is as follows:

Step1: sort data by value

Step2: set the subset of pairs of inter sectional clusters and adjacent clusters;

Step3: Put the most frequent cluster as the cluster tree root;

Step4: Put the intersect or adjacent clusters as the left of the parent;

Step5: The rightmost leaf of tree is the isolate cluster sets.

4 Experiment

We have executed the experiments using the same data sets by different algorithms. The experiment can illustrate the feasibility of the improved Kmeans.

<table>
<thead>
<tr>
<th>K</th>
<th>Time1</th>
<th>Time2</th>
<th>K</th>
<th>Time1</th>
<th>Time2</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>0.059079</td>
<td>0.7646</td>
<td>300</td>
<td>6.801531</td>
<td>64.150813</td>
</tr>
<tr>
<td>20</td>
<td>0.09199</td>
<td>1.161172</td>
<td>400</td>
<td>12.04738</td>
<td>73.476065</td>
</tr>
<tr>
<td>30</td>
<td>0.170177</td>
<td>1.878956</td>
<td>500</td>
<td>16.251014</td>
<td>103.036793</td>
</tr>
<tr>
<td>50</td>
<td>0.332438</td>
<td>2.679017</td>
<td>600</td>
<td>22.906389</td>
<td>116.773989</td>
</tr>
<tr>
<td>60</td>
<td>0.438395</td>
<td>3.390311</td>
<td>700</td>
<td>31.518395</td>
<td>121.329758</td>
</tr>
<tr>
<td>70</td>
<td>0.547304</td>
<td>3.633841</td>
<td>800</td>
<td>39.707395</td>
<td>167.421235</td>
</tr>
<tr>
<td>100</td>
<td>0.930863</td>
<td>4.778202</td>
<td>900</td>
<td>49.736557</td>
<td>173.411907</td>
</tr>
<tr>
<td>200</td>
<td>3.134636</td>
<td>5.966757</td>
<td>1000</td>
<td>60.819052</td>
<td>211.50758</td>
</tr>
</tbody>
</table>

We get the contrast result of running time though the traditional Kmeans algorithm and the improved Kmeans from Table 1. In Table 1, the time1 represents the executing time of Kmeans based on density constraint; the time2 shown running time of the traditional Kmeans algorithm. So, the improved Kmeans algorithm, which the paper has provided, can be seen as an available method for point clouds organization, especially, for the access data domain, and this is important to reverse reconstruction based on point cloud, for example feature extraction and data reduction.

In this paper, in order to adapt to the large-scale point cloud data processing, the traditional Kmeans algorithm has been improved. Among them, the data density is an important thing; it has been set as a computing condition of convergence, instead of the Euclidean distance. And it can find neighboring of data domain fast during
reconstruction. At last, the whole algorithm of point data organization based on clustering, has been proved to be feasible through the experiment.

Acknowledgments. The paper is funded by Science and Technology Department of Hebei Province in China (No.14K50120D), and it is sponsored by Hebei Education Department of Hebei Province in China (No. QN2014174); Top-notch young talent Foundation of Hebei Province of China under Grant (No. BJ2014023); National Natural Science Foundation of China (No.61440012).

References