Deep Learning Based Layered Point Cloud Compression for Representing Shape Aware Level of Detail

Hideaki Kimata

kimata@ieee.org
Kogakuin University, Shinjuku-ku, Tokyo, Japan
Keywords: point cloud compression, deep learning, level of detail

ABSTRACT

Point cloud compression methods based on deep learning are being studied. Extracting a part of point cloud from compressed data to grasp the characteristic shape of an object is an important function. I propose ranking and hierarchical coding that can sequentially decode main features of point cloud.

1 Introduction

Point cloud is expected to have a wide range of applications such as VR / AR and driving support. Research on compression coding of point cloud is underway. The international standardization organization MPEG is also promoting the standardization of G-PCC (Geometry based Point Cloud Compression) and V-PCC (Video based Point Cloud Compression), which use classical coding methods. Research on point cloud compression coding utilizing deep learning is progressing with the aim of achieving further compression efficiency.

Since a point cloud is a set of points, the required computational resource increases in proportion to the amount of points to be handled. Since it does not have the concept of standardized resolution like image, it requires unlimited computational resources in a sense. Therefore, when the point cloud is used in an actual application, a mechanism for extracting a desired number of point clouds from the compressed stream of the point cloud is required. In G-PCC, the so-called spatial scalable function is realized by expressing the point cloud with hierarchical voxels. The spatial scalable function makes it possible to decode and utilize a point cloud corresponding to a desired number of points, however, in the hierarchical voxel structure, the shape of the entire object represented by the point cloud is uniformly enlarged or reduced, so it is not suitable for partially grasping the detailed shape. Further, since the accuracy of the position of each point depends on the size of the voxel, there is a problem that the accuracy of the position is gradually lowered in the voxel having a low spatial resolution.

In the method studied based on deep learning, the point cloud is divided into blocks and compression coding is performed in units of blocks, similar to image coding. Since the point cloud can be decoded in block units, it is suitable for grasping the partial diameter. However, in order to grasp the shape of the entire point cloud, it is necessary to decode all the points.

Therefore, in this research, I propose a method suitable for grasping the shape of the entire point cloud while enabling partial decoding based on the deep learning-based method., The proposed method has a structure that can sequentially decode points necessary for grasping the overall or detail of shape, instead of having a spatially uniform hierarchical voxel structure.

2 Related Work

In V-PCC, the position of a point cloud is projected and expanded on a two-dimensional plane and encoded as an image. There is no functionality to extract a part of the point cloud in the three-dimensional space. In G-PCC, the position is represented by a three-dimensional voxel, multiple voxels with different spatial resolutions are prepared, and whether or not a point exists in the voxel is encoded. Based on hierarchical voxel structure, the voxels with high resolution are sequentially linked to the voxels with low resolution to form an octa-tree structure and encoded. Since voxels with low spatial resolution are also included in the coded data, it is suitable for grasping the shape of the entire point cloud. However, in order to grasp the partial shape of the point cloud, it is necessary to decode all related points over multiple resolutions. Decoding is performed from low spatial resolution voxels, followed by high resolution voxels in sequence [1].

In deep learning-based coding [2],[3], a point cloud is divided into blocks, and each block is compressed by a DNN (Deep Neural Network) to which an autoencoder is applied. Decoding is performed in block units. When optimizing rate distortion using lambdas, the models of

Fig.1 Examples of simplified polygon (Longdress)

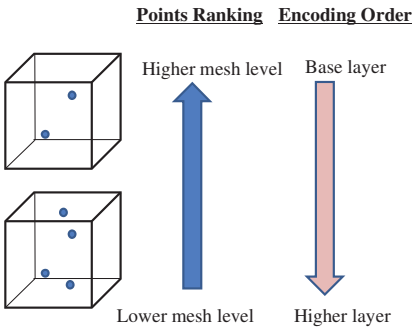


Fig.2 Relation of points rank and encoding order

the autoencoder are learned corresponding to each lambda. In addition, a method has been proposed in which a plurality of models are prepared for the autoencoder and adaptively switched and encoded [3]. Multiple models are trained according to the focal loss parameters. Then, when encoding, the trained model is selected and encoded depending on the three-dimensional distribution of voxels in the block. It is reported that a deep learning based solution outperforms G-PCC in compression efficiency.

3 Proposed method

3.1 Points ranking

The points in the point cloud are ranked so that the points can be sequentially decoded and the entire shape and detail of it can be grasped. To represent and evaluate the main features of the point cloud, a mesh is constructed from the point cloud. A point in a point cloud correspond to a vertex of a mesh. The edges and faces of a mesh are generated by the Alpha Shape [4].

Then, starting from the mesh generated from original point cloud, a series of mesh are recursively generated, which are used as a reference for ranking points in the point cloud. A mesh in the series is indicated by a mesh level. The higher level mesh is generated by degeneration of the lower level mesh so as to retain the main characteristics of the shape. The main feature here is heuristic. It is expected that important features of the shape of an object will be learned by the development of data analysis of point clouds such as PointNet [5], but at present, evaluation is performed only from a subjective point of view. Techniques for reducing the number of polygons while retaining the main features of the mesh are being studied. In the proposed method, the simplification method proposed by Garland et al. [6] is adopted to reduce the number of polygons. In this method, when a vertex pair is degenerated to a vertex, the total distance between a plurality of planes using the vertex is defined as an error, and the vertex is selected and degraded so that the error is minimized. By repeating simplification multiple times, a series of mesh are generated.

By repeating simplification, the number of polygons can be reduced step by step. Figure 1 shows an example when the number of polygons is reduced. It can be seen that the

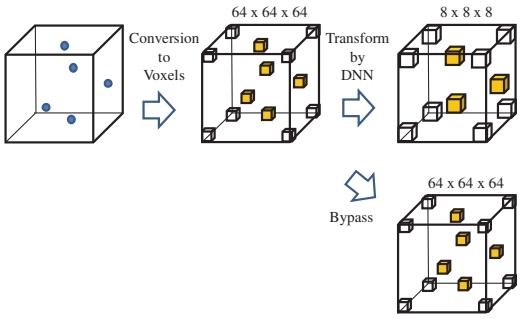


Fig.3 Diagram of encoding a block

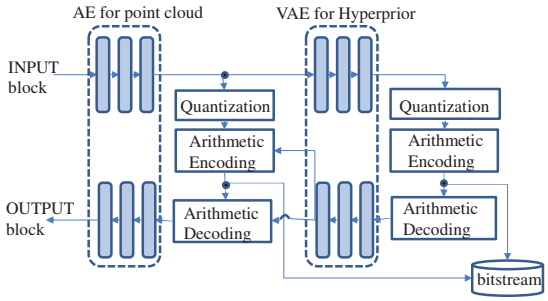


Fig.4 Diagram of encoder and decoder using DNN

number of polygons decreases while maintaining the main features. Since the simplification process degenerates the pair of vertices to one vertex, the vertex is generated at a position different from the vertex position of the original mesh. Therefore, for the result of each simplification, the nearest vertices in the original mesh are replaced. Then, the vertices of the obtained mesh are treated as a point cloud. From the mesh level, the points are ranked. By the above processing, attribute information about the rank is given to each point in the original point cloud.

3.2 Layered Point Cloud Compression

According to rank of the points, a layer structure of a point cloud is constructed. Since higher rank points have more main features, they are treated as more basic information. Therefore, the higher the rank of points, the lower the layer for coding, as depicted in Figure 2. Note that all the mesh levels are not treated as layers and a layer for encoding is selected from points of mesh levels. By having this layer structure, it becomes possible to sequentially decode the points of a layer and present detailed features sequentially from the main features. Further, since the coding is performed in block units, it is possible to decode only the desired portion in detail.

Figure 3 shows the overall diagram of the encoder. After dividing into blocks of appropriate size, the inside of the block is further divided into 64x64x64 voxels. Then, if there is a point in the voxel, set 1 in the voxel, and if there is no point in the voxel, set 0 in the voxel. Figure 4 shows the structure of the encoder and decoder that use deep learning. The DNN is similar to the method proposed by Guarda et al. [3]. The DNN transforms the

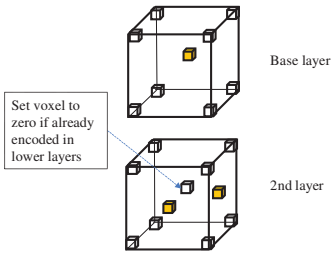


Fig.5 Relation of voxels in base and a high layer

1 or 0 pattern into a latent variable with 8x8x8 values. Then, the latent variable with 8x8x8 values is quantized and coded with arithmetic coding. Further variational autoencoder is applied to the latent variables to obtain the parameters that control this arithmetic coding. When the DNN learns this transformation or the inverse transformation, the focal loss is used for the loss function. Focal loss is an extension of binary cross entropy, as described by the following.

$$\mathrm{FL(u,v)} = \begin{cases} -\alpha(1-v)^{\gamma}\log v, & u = 1 \\ -(1-\alpha)v^{\gamma}\log(1-v), & u = 0 \end{cases}$$

where u is the original voxel binary value and v is the corresponding reconstructed voxel probability score. Binary cross entropy is used because it is treated as a classification problem. The focal loss is controlled by the parameter α. Parameter α is the key class imbalance. Larger values of α increase the importance of the '1' valued voxels. For training DNN, rate distortion performance is considered. Loss function is as follows,

$$\text{Loss Function} = \text{Distortion} + \lambda \times \text{Coding rate}$$

where Distortion is measured by the focal loss. One model for each parameter α is trained. In encoding a block, a model is selected according to rate distortion performance with the certain value of parameter λ.

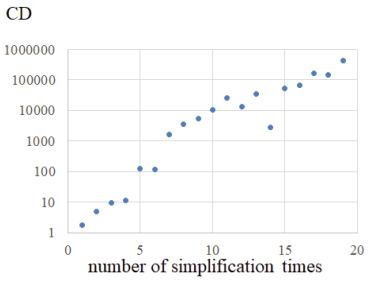
A method to reduce redundancy between layers in the process of encoding with a layer structure is also proposed. If the point is encoded in a lower layer, it is not necessary to encode the voxel information in the next and subsequent higher layers. This relation is shown in Figure 5. Any value of 0 or 1.can be set for such voxels. In this paper for evaluation, all associated values are set to 0.

And especially for encoding a block which has few voxels of 1, the bypass mode is proposed, as depicted in Fig. 3. In the bypass mode, voxel position of 1 in a block is quantized and encoded instead of encoding 1 or 0 voxel pattern.

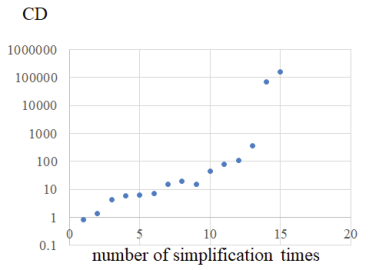
4 Evaluation and Discussion

4.1 Points Ranking

Simplification of a mesh was evaluated in terms of similarity to the original point cloud. In this paper, CD (Chamfer Distance) was used for evaluation metric, as it is



(a) House without roof



(b) Longdress

Fig.6 Results of CD

Fig.7 Comparison of proposed method and G-PCC (House without roof)

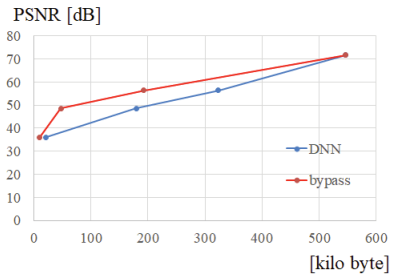
used for generative model of point clouds [7]. CD is defined as follows.

$$\mathrm{CD}(X,Y) = \sum_{x\in X}\min_{y\in Y}||x-y||_2^2 + \sum_{y\in Y}\min_{x\in X}||x-y||_2^2$$

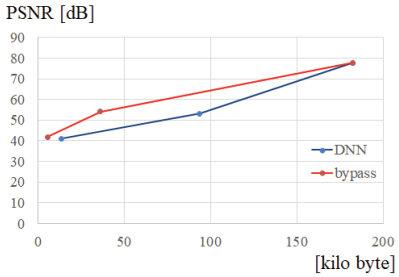
where X and Y are two point clouds with the same number of points.

To evaluate by CD, points are sampled from a mesh. Figure 6 shows CD calculated from original point cloud for mesh levels. Test point cloud data "House without roof" and "Longdress", which are prepared for MPEG PCC standardization, were used. "House without roof" has 5,001,077 points and represents walls of house and a tree. "Longdress" has 857,966 points and represents a dressed human shape. We can see almost proportional relation between mesh level and logarithm of CD.

Shape of ranked points generated by simplification is compared with spatial scalable layered structure realized by G-PCC. Figure 7 shows the results where the almost same number of points exist in both simplifications. For proposed simplification, mesh level is 10th. The mesh was generated from points by Alpha Shape. It can be seen that the proposed method expresses the dents in



(a) House without roof



(b) Longdress

Fig.8 Rate distortion performance of layered coding

the walls of the house more apparently than G-PCC.

4.2 Layered Coding

Compression performance is evaluated. The PSNR calculation method used in MPEG PCC international standardization [8] was used. PSNR was calculated for the original point cloud. For point clouds, PSNR of the point-to-point distortion means the symmetric error between every point in the original point cloud and its closest neighbor in the decoded point cloud. The DNN models were trained with the parameter α set to 0.5, 0.6, 0.7, 0.8 or 0.9 and with the parameter γ set to 2 in the focal loss.

Figure 8 shows the amount of code and the result of PSNR. The blue dots and lines are when each block is coded using DNN, and the red dots and lines are when the blocks are coded in bypass mode for encoding high layer. Three high layers for “House without roof” (corresponding to 10th, 6th, 4th mesh level) and two high layers for “Longdress” (corresponding to 8th, 4th mesh level) were encoded in addition to the base layer. The parameter λ was set to 500. From this figure, it can be seen that deep learning solution works efficiently for layered coding. In addition, the bypass mode has higher performance in terms of rate distortion. At higher layers, points are sparsely present for the structure of the block, so bypass mode is thought to have contributed effectively.

DNN was compared with bypass mode in encoding in several rate distortion relation. Figure 9 shows the results for a high layer in a two-layer structure (corresponding to 4th mesh level), in which the parameter λ was either 20000,5000,1500,900 or 500. The results by DNN encoding “House without roof” are indicated by blue dots and lines and the result of encoding in bypass mode is indicated by red dot. It can be seen that DNN can decrease

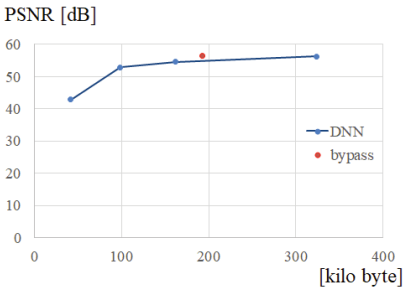


Fig.9 Comparison of DNN and bypass mode

the coding rate than bypass mode. Therefore it was found that DNN and bypass mode should be properly selected in terms of rate distortion optimization.

5 Conclusion

I propose points ranking and hierarchical coding method of point cloud that can sequentially decode main features of the shape of the object. The proposed method extends the state-of-the-art deep learning based method for encoding, and it is demonstrated that the proposed method achieves higher rate distortion performance.

Acknowledgment

This work was supported by JSPS KAKENHI Grant Number JP22K12098.

References

[1] S. Schwarz et al., "Emerging MPEG Standards for Point Cloud Compression," IEEE Journal on Emerging and Selected Topics in Circuits and Systems, Vol.9, No.1, pp. 133-148, (Mar. 2019).

[2] M. Quach, G. Valenzise, F. Dufaux, “Improved Deep Point Cloud Geometry Compression,” Proc. IEEE International Workshop on Multimedia Signal Processing (MMSP’2020), (Sep. 2020).

[3] A. F. R. Guarda, N. M. M. Rodrigues and F. Pereira, "Adaptive Deep Learning-Based Point Cloud Geometry Coding," IEEE Journal of Selected Topics in Signal Processing, Vol. 15, No. 2, pp. 415-430, (Feb. 2021).

[4] H. Edelsbrunner and E. P. Mücke, “Three-dimensional alpha shapes,” ACM Trans. Graph. Vol. 13, Issue 1, pp. 43–72 (Jan.1994).

[5] C. R Qi, H. Su, K. Mo, and L. J Guibas, “Pointnet: Deep learning on point sets for 3d classification and segmentation,” Proc. CVPR, (Jul. 2017).

[6] M. Garland and P. S. Heckbert, “Surface simplification using quadric error metrics,“ Proc. ACM SIGGRAPH '97. pp.209–216 (Aug. 1997).

[7] P. Achlioptas, O. Diamanti, I. Mitliagkas, and L. Guibas, “Learning representations and generative models for 3d point clouds,” Proc. ICML, (Jul. 2018).

[8] ISO/IEC JTC1/SC29/WG11 N19084. Common Test Conditions for Point Cloud Compression. Brussels, Belgium, (Jan. 2020).