

REFINING THE BOUNDING VOLUMES FOR LOSSLESS COMPRESSION OF VOXELIZED POINT CLOUDS GEOMETRY

Emre Can Kaya* Sebastian Schwarz† Ioan Tabus*

* Tampere University, Tampere, Finland

† Nokia Technologies, Munich, Germany

ABSTRACT

This paper describes a novel lossless compression method for point cloud geometry, building on a recent lossy compression method that aimed at reconstructing only the bounding volume of a point cloud. The proposed scheme starts by partially reconstructing the geometry from the two depthmaps associated to a single projection direction. The partial reconstruction obtained from the depthmaps is completed to a full reconstruction of the point cloud by sweeping section by section along one direction and encoding the points which were not contained in the two depthmaps. The main ingredient is a list-based encoding of the inner points (situated inside the feasible regions) by a novel arithmetic three dimensional context coding procedure that efficiently utilizes rotational invariances present in the input data. State-of-the-art bits-per-voxel results are obtained on benchmark datasets.

Index Terms— Point Cloud Compression, Context Coding, Lossless Compression, Arithmetic Coding

1. INTRODUCTION

The lossless compression of point clouds was thoroughly studied and is currently under standardization under MPEG [1],[2] and JPEG activities [3]. The research literature on point cloud compression includes a lot of methods based on octree representations, e.g., [4, 5, 6, 7, 8]. The current GPCC lossless method [9] is also based on octree representations [10], where a point cloud is represented as an octree, which can be parsed from the root to the leaves and at each depth level in the tree, one obtains a lossy reconstruction of the point cloud at a certain resolution, while the lossless reconstruction is obtained at the final resolution level in the octree. At each resolution level, an octree node corresponds to a cube at a particular 3D location and the octree node is labeled by one if within that cube there is at least one final point of the point cloud. By traversing in a breadth-first way, one can obtain a lossy reconstruction at each depth of the tree, while the lossless reconstruction is obtained at the final resolution. Octree-based approach is attractive for providing a progressive-in-resolution reconstruction.

Recently, good lossless performance was achieved by methods based on successive decomposition of the point cloud data into binary images [11, 12], without using octree representations.

This paper describes a novel lossless point cloud compression algorithm which is based on Bounding Volumes [13] for G-PCC. In the Bounding Volumes, all reconstructed points truly belonged to the point cloud, but some points, specifically inner points in the transversal sections of the point cloud, were not encoded at all. In this work, a complete lossless process, overlapping in the first stage with the Bounding Volumes method (encoding a front and a back projection), but diverging from the previous method, already at the second stage, that of encoding the true boundary of the feasible region and making it less restrictive, getting rid of the requirement of decomposing the point cloud into tubes having single connected component sections.

Compared to the lossy Bounding Volume method, the newly introduced encoding process includes an extensive context coding scheme that can finally provide a lossless reconstruction of the point cloud. This novel scheme is based on carefully selected context coding methods and intensively uses exclusion to avoid unnecessary testing for the occupancy of the locations which are already known.

2. PROPOSED METHOD

Consider a point cloud having the resolutions N_x , N_y , N_z along the axes x , y , z , respectively. The points are encoded in two stages. In Stage I, two projections of the points cloud are encoded; the front and the back projections on xy plane. These projections are two depthmap images, each with $N_x \times N_y$ pixels. Then, the coding proceeds along the Oy axis of the 3D coordinate system, drawing transverse sections of the point cloud parallel to the zOx plane and encoding each such section in Stage II. An overview of the method is presented on Fig. 1. The regularity of the geometric shapes, including smoothness of the surface and the continuity of the edges, are exploited by using context models, where the probability of occupancy of a location is determined by the occupancy of its neighbor locations in 3D space, by including the causal (previously encoded) 3D neighbors from the

current and the past sections.

2.1. Stage I: Encoding a front and a back depthmap projection

The first encoding stage is intended for defining and encoding the maximal and minimal depthmaps (representing heights above the Oxy plane) resulting in exclusion sets containing large parts of the space. The minimal depthmap, Z_{min} , has at the pixel (x, y) the value, $Z_{min}(x, y)$ equal to the minimum z for which (x, y, z) is a point in the original point cloud. Similarly, the maximal depthmap, Z_{max} has at the pixel (x, y) the value, $Z_{max}(x, y)$ equal to the maximum z for which (x, y, z) is a point in the original point cloud. If no point with (x, y) exists in the point cloud, then it is set $Z_{min}(x, y) = Z_{max}(x, y) = 0$. The encoding of these depthmaps is performed by CERV [14], which encodes the geodesic lines using contexts very rich in information.

2.2. Stage II: Encoding the remaining points

At Stage II, we sweep the point cloud along the y dimension, stepping y_0 from 0 to $N_y - 1$, and we reconstruct all the points in the current section y_0 in the $(N_z \times N_x)$ binary image R (current reconstruction). At every y_0 , points projected to the depthmaps are already known to the decoder and we initialize the current reconstruction R with the projected points such that, $R(Z_{max}(x, y_0), x) = 1$ and $R(Z_{min}(x, y_0), x) = 1$ for every $Z_{max}(x, y_0) > 0$ and $Z_{min}(x, y_0) > 0$. R will be updated whenever a new point is encoded or decoded. We note that, in the binary image R , we know at each column x which are the lowest and the highest occupied points (minimal and maximal depths). We consequently construct the binary image F of feasible regions, i.e., of locations in the plane that are possibly occupied (the magenta pixels in Fig. 2(b)). Formally, $F(z, x) = 1$ for all (z, x) pair satisfying $Z_{min}(x, y_0) \leq z \leq Z_{max}(x, y_0)$. Using R and F , we initialize a binary image K , where 1 denotes that the occupancy of a location is known. For example, the locations outside the feasible region are known to be unoccupied, hence, $K(z, x) = 1$ for those locations. The true points in the section, that we need to losslessly encode, are marked in a binary image denoted T (see Fig. 2(a)), and the reconstructed points in the past section (at $y_0 - 1$, that is already known to the decoder) are marked in a binary image P .

In the image R the already reconstructed true points belonging to depthmaps form a set ϕ of pixels. We perform a morphological dilation of the set ϕ using as structural element the 3×3 element. This obtained set of locations is traversed along the rows of the 2D image and is stored in a list L . We also initialize a binary marker image M to mark the pixels already in the list L . After this initialization step, the list L is processed sequentially, starting from the top of the list, processing a current pixel (z, x) . Both encoder and decoder check whether $K(z, x) = 1$, and if yes, the point is

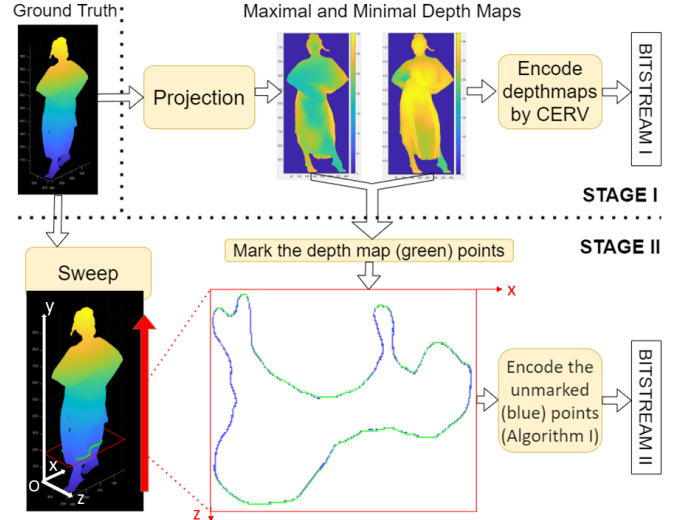


Fig. 1. Overview of the proposed method: In Stage I, minimal and maximal depthmaps Z_{min} and Z_{max} are generated by projecting the point cloud along z axis. The depthmaps are encoded by CERV[14]. In Stage II, the point cloud is swept through y axis. The points that are not already encoded by CERV (blue points) are encoded as described in Sections 2.2 and 2.3.

removed from the list, since its occupancy status is already known. Otherwise, if $K(z, x) = 0$, we transmit the value of $T(z, x)$ using arithmetic coding with the coding distribution stored at the context ζ . After that, the counts of symbols associated with the context ζ are updated. K is updated as $K(z, x) \leftarrow 1$, and the reconstructed image is updated as $R(z, x) \leftarrow T(z, x)$. If the value $T(z, x) = 1$, we include to the list any neighbor (z_n, x_n) of (z, x) , (in 8-connectivity) for which $K(z_n, x_n) = 0$ and for which the marked status is 0, $M(z_n, x_n) = 0$. After inclusion, the marked status is set to 1, $M(z_n, x_n) = 1$. This procedure is repeated until the list becomes empty. At the end, we have encoded all the points that are connected to the boundary of the feasible region by a path of connected pixels (in 8-connectivity). After the final section $y_0 = N_y - 1$ is processed, all the points that are connected to the points contained in the initial two depthmap images are encoded. For the voxelized point clouds, this outer shell of points contains the majority of the points in the point cloud. The remaining points (if any) are encoded by processing additional shells, as described in Section 2.4.

2.3. Normalized Contexts

One of the most efficient ways of utilizing the contextual information that is needed in Stage II is described here. In order to show the elements forming the context, it is illustrated in Fig. 2(c) the ground truth occupancies for the cur-

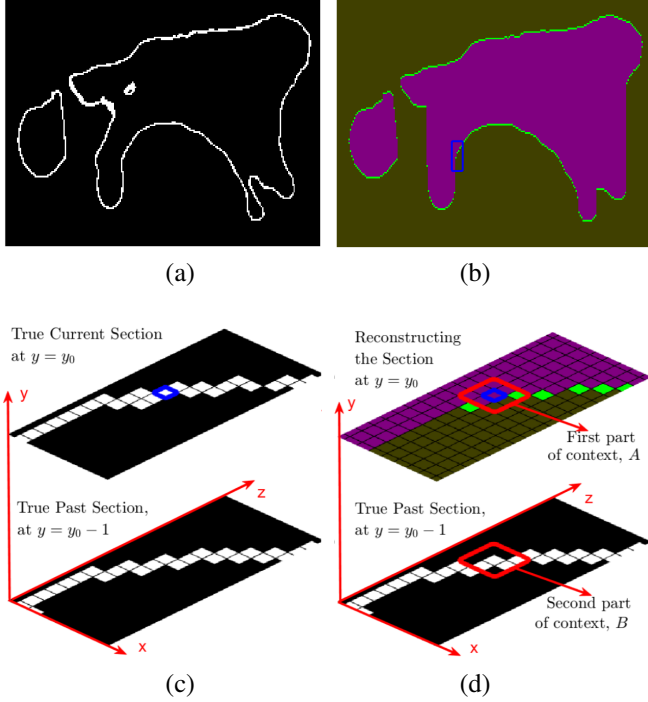


Fig. 2. (a) True current section; (b) The feasible region (magenta), points from the depth maps (green), forbidden region (khaki); (c) Details of the ground truth inside the blue rectangle from (b); (d) Selection of the two part context for encoding the blue marked pixel

rent section ($y = y_0$) and the past section ($y = y_0 - 1$), showing in white the (true) occupied pixels in these sections. On Fig. 2(d), we show the same area during the reconstruction process, which advances section by section such that, at the moment of reconstructing the section $y = y_0$, the section $y = y_0 - 1$ is fully known, and the second part of the context, called matrix B , can be extracted and contains the true occupancy status at section $y = y_0 - 1$. The context A from the current section for every candidate pixel (z, x) is extracted from the 3×3 neighbourhood of the pixel (i.e., the pixels in the red square).

When encoding the blue marked pixel, the pixels considered as the 3×3 A matrix part of context are those surrounded by the red contour. Each pixel might be green (already encoded in Stage I), for which the status is known and occupied ($K(z, x) = 1$ and $R(z, x) = 1$), khaki for forbidden pixels with status known and unoccupied ($K(z, x) = 1$ and $R(z, x) = 0$), and finally magenta for feasible i.e., not yet known ($K(z, x) = 0$). The context extracted from the current section forms the matrix A and that from the past section forms the matrix B (which are later rotated for normalizing and are combined to form the final contexts).

The procedure for encoding the points at a section $y = y_0$ with normalized contexts is summarized in Algorithm 1. Con-

Algorithm 1 Stage II of encoding

Require: T : True section binary image at $y = y_0$ ($N_z \times N_x$)
 P : True section binary image at $y = y_0 - 1$ ($N_z \times N_x$)
 R : Current reconstruction at $y = y_0 - 1$ ($N_z \times N_x$)
 F : Feasible Regions bin. image derived from R (Sect. 2.2)
 K : Binary image of known locations $K \leftarrow \bar{F} + R$
 L : Pixels to be processed $L \leftarrow \{(z, x) \ni R(z, x) = 1\}$
 M : Binary image of pixels that has been to L . $M \leftarrow R$
while $L \neq \emptyset$ **do**
 Read (z, x) from the top of L
 Extract a 3×3 matrix $R_{3 \times 3}$ from R centered at (z, x)
 Extract a 3×3 matrix B from P centered at (z, x)
 Extract a 3×3 matrix $K_{3 \times 3}$ by cropping K around (z, x)
 $A \leftarrow R_{3 \times 3} + K_{3 \times 3}$
 Find normalizing rotation $\alpha^*(A)$ and form A_{α^*}
 Use $\alpha^*(A)$ to rotate B as B_{α^*}
 Form the context $\zeta = (I(A_{\alpha^*}), J(B_{\alpha^*}))$
 Encode $T(z, x)$ using the context ζ
 if $T(z, x) == 1$ **then**
 Append to L every neighbor (z_n, x_n) of (z, x) for which $K(z_n, x_n) = 0$ and $M(z_n, x_n) = 0$
 If (z_n, x_n) is appended, $M(z_n, x_n) \leftarrow 1$
 end if
 Update R : $R(z, x) \leftarrow T(z, x)$
 Update K : $K(z, x) \leftarrow 1$
 Remove (z, x) from L
end while

sider that we need to encode $T(z, x)$. The first part of the context uses the values of the already reconstructed pixels that are 8-neighbors of (z, x) and also the information about which of the pixels were already known. The values of the pixels in the ternary image $R_k = R + K$ have the following significance: $R_k(z, x) = 0$ if the value of $T(z, x)$ is not known yet; $R_k(z, x) = 1$ if the value of $T(z, x)$ was encoded and $T(z, x) = 0$; $R_k(z, x) = 2$ if the value of $T(z, x)$ was encoded, and $T(z, x) = 1$. We consider first the 3×3 square, centered at (z, x) , cropped from the image R_k , and we denote it as a 3×3 matrix A . The elements of A belong by construction to 0, 1, 2. The second part of the context is the 3×3 binary matrix B formed from P at (z, x) . The information from A and B is used to form the context. For example scanning by columns we get a one-to-one correspondence between A and $I(A) = \sum_{j=0}^2 \sum_{i=0}^2 A_{ij} 3^{i+3j}$. Similarly there is a one-to-one correspondence between B and $J(B) = \sum_{j=0}^2 \sum_{i=0}^2 B_{ij} 2^{i+3j}$. We combine them to a context label $\zeta = (I(A), J(B))$.

The context information in A and B is further normalized in the following way: We consider performing context collapsing operations, such that if we would perform a rotation by $\alpha \in \{0, \pi/2, \pi, 3\pi/2\}$ of each of the images R , T and K around the pixel (z, x) , the value of the resulting normalized context is the same. We consider first the 3×3

Table 1. Average Rate for the first 200 frames from MVUB [15] and 8i [16] datasets for the proposed Bounding Volume Lossless (BVL) encoder compared to recent codecs.

Sequence	Average Rate [bpv]			
	P(PNI)[8]	TMC13[9]	DD[12]	BVL
Microsoft Voxelized Upper Bodies [15]				
Andrew9	1.83	1.14	1.12	1.17
David9	1.77	1.08	1.06	1.10
Phil9	1.88	1.18	1.14	1.20
Ricardo9	1.79	1.08	1.03	1.05
Sarah9	1.79	1.07	1.07	1.08
Average	1.81	1.11	1.08	1.12
8i Voxelized Full Bodies [16]				
Longdress	1.75	1.03	0.95	0.91
Loot	1.69	0.97	0.91	0.88
Redandblack	1.84	1.11	1.03	1.03
Soldier	1.76	1.04	0.96	0.96
Average	1.76	1.04	0.96	0.94

matrix A . Apply the α rotation around the middle pixel and denote A_α the resulting 3×3 matrix. Compute for each of $\alpha \in \{0, \pi/2, \pi, 3\pi/2\}$, the matrix A_α and the weighted score of it $W(A_\alpha)$ and pick as canonical rotation that α_* for which the weighted score $W(A_{\alpha_*})$ is the largest. Hence, the four rotated matrices A_α with $\alpha \in \{0, \pi/2, \pi, 3\pi/2\}$ will be represented only by A_{α_*} . This process of collapsing the four matrices into a single one is performed offline once, resulting in a tabulated mapping $\alpha_* \leftrightarrow A$ and another mapping $I_* \leftrightarrow A$, which realize the mappings of the context to the canonical one, stored in look-up tables. As an example of the weighting score $W(A)$, we consider the vector $v = [A_{00}A_{01}A_{10}A_{02}A_{11}A_{20}A_{12}A_{21}A_{22}]$ and form $W(A) = \sum_{k=0}^8 v_k 3^k$, giving in this way a larger weight to those elements which are close to the corner $(0, 0)$ of A . The normalized context is found in the following way: At each point (z, x) the matrix A is formed from $R+K$ and the canonical rotation index α_* for this matrix is computed. Also the corresponding rotated matrix A_0 is computed. The second part of the context is the 3×3 matrix B formed from P around (z, x) . The matrix B is rotated by the previously determined α_* around its center, yielding a matrix B_0 . Now the context to be used for encoding $T(z, x)$ is constructed from A_0 and B_0 as context $\zeta = (I(A_0), J(B_0))$.

2.4. Repetitive peeling-off process for complete reconstruction of more complex point clouds

After Stage II, the reconstruction contains all the points forming the outer surfaces of the point cloud and all the inner points connected to these outer surface points, i.e., all points

that are connected by a path in 3D space (in 26-voxel connectivity), to the initial points recovered in Stage I from the two depthmap images. However, there are complex point clouds, for example those representing a building and objects inside, where some objects are not connected by a 3D path to the outermost points. In that case one can repeat the encoding process shell by shell, in a peeling-off operation, where we encode first the outermost shell, defined by the points represented in the maximal and minimal depthmaps and all points connected to these points, and then we reapply the same process to the remaining un-encoded points. If needed, this peeling-off process can be applied several times. In this work, there are maximally 2 shells peeled-off and the remaining points (if any) are simply written in binary representation into the bitstream.

3. EXPERIMENTAL WORK

The algorithm is implemented in C and the experiments were carried out on two point cloud datasets namely, 8i Voxelized Full Bodies [16] and Microsoft Voxelized Upper Bodies [15]. The average bits per occupied voxel results (average rates) are presented on Table 1. For each point cloud, all 6 possible permutations of the 3 dimensions are tried and the best rate obtained is kept and reported here. It is observed that, proposed method performs better than the other methods on the 8i dataset. On the other hand, on MVUB dataset, our results are slightly worse than TMC13 [9] and Dyadic Decomposition [12]. Additionally, we test BVL and TMC13 on all the point clouds from the Cat1A MPEG Database [17] having original resolutions of 10 and 11 bits. These were quantized to 10, 9 and 8 bits as well to test the performance in lower resolutions. BVL outperformed TMC13 in average at 10 bits by 6.6%, at 9 bits by 5.2%, at 8 bits by 2%. At 11 bits, TMC13 outperformed BVL by 1.8%. For all of the point clouds mentioned in this work, the decoding resulted in a perfect lossless reconstruction.

Encoding and decoding durations for BVL were measured to be 7.3 sec and 7.8 sec, respectively. On the same machine, encoding with TMC13 took 1.1 sec. All durations are measured on a single frame of the 10 bits longdress sequence by running the algorithm 10 times and taking the median. While the durations are not competitive with TMC13, it should be noted that the execution time is not yet carefully optimized.

4. CONCLUSIONS

We proposed a lossless compression method where the first stage is constructing a bounding volume for the point cloud and the following steps succeed at adding all the remaining points at a competitive bitrate, achieving state-of-the-art results for the full body datasets, and comparable results to the current GPCC standard on the upper body dynamic datasets.

5. REFERENCES

- [1] S. Schwarz, M. Preda, V. Baroncini, M. Budagavi, P. Cesar, P. A. Chou, R. A. Cohen, M. Krivokuća, S. Lasserre, Z. Li, 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, 2018.
- [2] L. Cui, R. Mekuria, M. Preda, and Eu. S. Jang, “Point-cloud compression: Moving picture experts group’s new standard in 2020,” *IEEE Consumer Electronics Magazine*, vol. 8, no. 4, pp. 17–21, 2019.
- [3] T. Ebrahimi, S. Foessel, F. Pereira, and P. Schelkens, “JPEG Pleno: Toward an efficient representation of visual reality,” *Ieee Multimedia*, vol. 23, no. 4, pp. 14–20, 2016.
- [4] R. L. De Queiroz and P. A. Chou, “Compression of 3d point clouds using a region-adaptive hierarchical transform,” *IEEE Transactions on Image Processing*, vol. 25, no. 8, pp. 3947–3956, 2016.
- [5] R. Mekuria, K. Blom, and P. Cesar, “Design, implementation, and evaluation of a point cloud codec for tele-immersive video,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 4, pp. 828–842, 2016.
- [6] S. Milani, “Fast point cloud compression via reversible cellular automata block transform,” in *2017 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2017, pp. 4013–4017.
- [7] Diogo C Garcia and Ricardo L de Queiroz, “Intra-frame context-based octree coding for point-cloud geometry,” in *2018 25th IEEE International Conference on Image Processing (ICIP)*. IEEE, 2018, pp. 1807–1811.
- [8] D. C. Garcia, T. A. Fonseca, R. U. Ferreira, and R. L. de Queiroz, “Geometry coding for dynamic voxelized point clouds using octrees and multiple contexts,” *IEEE Transactions on Image Processing*, vol. 29, pp. 313–322, 2019.
- [9] “MPEG Group TMC13,” <https://github.com/MPEGGroup/mpeg-pcc-tmc13>, Accessed: 2020-03-20.
- [10] D. Meagher, “Geometric modeling using octree encoding,” *Computer graphics and image processing*, vol. 19, no. 2, pp. 129–147, 1982.
- [11] R. Rosário and E. Peixoto, “Intra-frame compression of point cloud geometry using boolean decomposition,” in *2019 IEEE Visual Communications and Image Processing (VCIP)*. IEEE, 2019, pp. 1–4.
- [12] E. Peixoto, “Intra-frame compression of point cloud geometry using dyadic decomposition,” *IEEE Signal Processing Letters*, vol. 27, pp. 246–250, 2020.
- [13] I. Tabus, E. C. Kaya, and S. Schwarz, “Successive refinement of bounding volumes for point cloud coding,” in *2020 IEEE 22nd International Workshop on Multimedia Signal Processing (MMSP)*. IEEE, 2020, pp. 1–6.
- [14] I. Tabus, I. Schiopu, and J. Astola, “Context coding of depth map images under the piecewise-constant image model representation,” *IEEE Transactions on Image Processing*, vol. 22, no. 11, pp. 4195–4210, 2013.
- [15] C. Loop, Q. Cai, S. O. Escolano, and P. A. Chou, “Microsoft voxelized upper bodies - a voxelized point cloud dataset,” *ISO/IEC JTC1/SC29 Joint WG11/WG1 (MPEG/JPEG) input document m38673/M72012*, 2016.
- [16] E. d’Eon, B. Harrison, T. Myers, and P. A. Chou, “8i voxelized full bodies - a voxelized point cloud dataset,” *ISO/IEC JTC1/SC29 Joint WG11/WG1 (MPEG/JPEG) input document WG11M40059/WG1M74006*, 2017.
- [17] S. Schwarz, G. Martin-Cocher, D. Flynn, and M. Budagavi, “Common test conditions for point cloud compression,” *Document ISO/IEC JTC1/SC29/WG11 w17766, Ljubljana, Slovenia*, 2018.