SPARSE REPRESENTATION FOR COLORS OF 3D POINT CLOUD VIA VIRTUAL ADAPTIVE SAMPLING

Junhui Hou[§]*, Lap-Pui Chau*[†]*, Ying He*[†]*, and Philip A. Chou*[‡]

[§]Department of Computer Science, City University of Hong Kong, Kowloon, Hong Kong
[†]Nanyang Technological University, Singapore 639798; [‡]Microsoft Research, Redmond, WA 98075
jh.hou@cityu.edu.hk, {elpchau, yhe}@ntu.edu.sg, pachou@ieee.org

ABSTRACT

Sparse signal representation has proven to be an extremely powerful tool in a wide range of engineering applications. However, most of the existing techniques are designed for regular data (such as audio signals and images/videos) that uniformly lies in regular Euclidian spaces. This paper aims at extending sparse representation for irregular data (such as colors of 3D point clouds) that is defined on irregular domains embedded in Euclidean spaces. Dealing with the irregular structure of such data via a virtual adaptive sampling process, we formulate sparse representation as an ℓ_0-norm regularized optimization problem. Experimental results show that the proposed algorithm outperforms the state-of-the-art algorithm to a large extent: with the same number of nonzero coefficients, we improve the reconstruction quality up to 5 dB; conversely, fixing the reconstruction quality, our method uses only 55% coefficients. Using compressive sensing theory, we provide an intuitive explanation on how and why our algorithm works well in practice.

Index Terms— 3D point cloud, voxelization, sparse representation, compressive sensing, compression, reconstruction

1. INTRODUCTION

A 3D point cloud consists of a set of 3D coordinates indicating the locations of points, along with one or more attributes (e.g., colors and normals) associated with each point, which can be used for representing 3D objects and scenes. Recent developments in 3D acquisition (e.g., computer vision, structured light and time of flight based depth sensors, etc.) make it relatively easy to obtain highly detailed 3D point clouds. This kind of data is becoming popular in emerging applications such as augmented reality, gaming, 3D telepresence, and immersive communications, since it allows free-view point rendering, adapts to represent objects of complex topology, and is computationally efficient. In spite of its popularity, some fundamental issues still exist, diminishing the use of such data. For example, it is common that a point cloud contains millions of points, leading to huge amounts of data, so effective and efficient compression schemes have to be developed due to limited network bandwidth and storage space [1], [2]. The acquired data may be defective due to occlusion or other factors (e.g., noise and holes), and thus, preprocessing operations have to be performed to restore it [3].

As in [2, 4, 5, 6, 7], we adopt the voxel based representation for unstructured 3D point clouds, that is, with a given stepsize 3D coordinates are quantized to regular and axis-aligned 3D grids of dimensions $2^L \times 2^L \times 2^L$ where L is the level. A voxel is said *occupied*, if it contains at least one point. The geometry of a voxel is an unsigned integer triple $\mathbf{v} \in \mathbb{R}^{3\times 1}$ corresponding to the 3D coordinate of the voxel corner, and the attributes, the average value of those of included points. An *unoccupied* voxel is transparent and devoid of other properties. Voxelized 3D point clouds can be efficiently organized and encoded using an octree structure [8]. Besides, fast techniques for producing and rendering such data have been developed [9].

Sparse representation (SR) has proven to be an extremely powerful tool for acquiring, representing and compressing high-dimensional signals. With SR, a signal is written as a linear combination of only a few atoms from a pre-specified basis or dictionary. The sparsity principle plays an important role in data modeling that is a crucial step for performing various operations such as restoration, compression, or for solving inverse problems. Therefore, techniques exploiting the sparsity of signals in a transform domain or dictionary have been popular in signal processing, ranging from the Fourier transform, discrete cosine transform (DCT), and wavelet transform to redundant dictionaries [10]. In particular, SR over a redundant dictionary has shown promise in various applications [11], such as denoising, classification, super resolution, restoration, and compression, just to name a few.

Motivation: It is natural and highly desirable to bring the recent advances in SR techniques to 3D point clouds to address the above-mentioned issues and others. However, it is not a straightforward extension since heretofore these algorithms have worked only with signals that are uniformly sampled in regular Euclidean spaces (e.g., audios, images, and

videos), while 3D point cloud data does not exhibit such characteristic [12]. Specifically, taking images as an example, since pixels are *uniformly* distributed in a regular 2D grid, with a predefined patch size $\sqrt{p} \times \sqrt{p}$ arbitrarily extracted patches can be respectively reorganized as vectorial signals in $\mathbb{R}^{p\times 1}$, and then sparsely coded over a basis/redundant dictionary of dimension $\mathbb{R}^{p\times 1}$. However, for a voxelized 3D point cloud, although the overall voxel set lies in a regular 3D grid, the set of occupied voxels is *non-uniformly* distributed in the space. As a result, the dimensions of vectorial signals defined on the occupied voxels in every $k \times k \times k$ block of voxels vary from block to block. Thus, the SR for 3D point clouds is more challenging.

In this paper we focus on attributes on 3D point clouds, such as colors and normals. Without loss of generality, we use color attributes (in RGB or YUV color spaces) as an example. We propose a very effective method to represent sparsely the colors on voxelized 3D point clouds. We employ a virtual adaptive sampling process to deal with the irregular structure so that the task can be elegantly formulated as an ℓ_0-norm regularized optimization problem, i.e., pursuit of the sparse coefficients with respect to an overcomplete dictionary. Experimental results demonstrate its effectiveness and superiority over the state-of-the-art method. We believe such an effective algorithm can contribute to 3D point cloud compression and other processing like denoising and restoration.

The rest of this paper is organized as follows. Section 2 reviews several existing algorithms followed by the proposed algorithm in Section 3. Experimental results are shown in Section 4. Section 5 provides an intuitive explanation to the proposed algorithm based on compressive sensing theory. Finally, Section 6 concludes this paper and points out directions for future work.

2. RELATED WORK

There are several methods proposed to obtain approximately sparse coefficients of colors on voxelized 3D point clouds in the transform domain for compression purposes. Zhang *et al*. [4] applied the graph transform (GT) to decorrelate such data; that is, a graph was formed for the occupied voxels within a 3D block, and the graph Laplacian matrix was obtained via the inverse distance (ID) model, whose eigenvector matrix was further used as the transform matrix. Queiroz and Chou [1] proposed a region-adaptive hierarchical transform (RAHT), which is a hierarchical sub-band transform that resembles an adaptive variation of a Haar wavelet. RAHT is more computationally efficient than GT while achieving comparable performance to GT. Cohen *et al*. [5] extended the well-known shape adaptive DCT (SA-DCT) [13] designed to code arbitrarily shaped regions in images to voxelized 3D point clouds. Note that the bases of these four algorithms are orthogonal.

3. THE PROPOSED ALGORITHM

As discussed in Section 1, the major technical challenge of SR-based modeling of a voxelized 3D point cloud arises from its irregularity (or non-uniformity): for a certain 3D block of size $k \times k \times k$, some voxels are occupied, while the others are empty; moreover, the number of occupied voxels are spatial varying. Such an irregular characteristic can be viewed as a virtual adaptive sampling process, i.e.,

$$\mathbf{x}_i = \mathbf{S}_i \mathbf{y}_i, \tag{1}$$

where $\mathbf{x}_i \in \mathbb{R}^{n_i\times 1}$ consists of colors[1] of occupied voxels of the i-th block, $\mathbf{y}_i \in \mathbb{R}^{k^3\times 1}$ is a virtual signal, containing colors of all voxels under the assumption that all voxels of the i-th block are occupied, $\mathbf{S}_i \in \mathbb{R}^{n_i\times k^3}$ ($n_i \ll k^3$) is a downsampling matrix, i.e., the identity matrix with reduced rows corresponding to unoccupied voxels of the i-th block, and n_i is the number of occupied voxels in the i-th block. Generally, colors within a small 3D block change little, i.e., having a locally smooth characteristic, indicating that $\mathbf{y}_i$ is a compressible signal. Thus, we further assume that $\mathbf{y}_i$ can be transformed into approximately sparse coefficients $\mathbf{c}_i \in \mathbb{R}^{k^3\times 1}$ by a full-rank basis $\mathbf{\Phi} \in \mathbb{R}^{k^3\times k^3}$, i.e.,

$$\mathbf{y}_i = \mathbf{\Phi}\mathbf{c}_i. \tag{2}$$

Furthermore, we formulate an optimization problem to recover the coefficients:

$$\min_{\{\mathbf{c}_i\}} \sum_{i=1}^{N} \|\mathbf{c}_i\|_0 \ \text{ subject to } \ \frac{1}{N}\sum_{i=1}^{N} \|\mathbf{x}_i - \mathbf{S}_i\mathbf{\Phi}\mathbf{c}_i\|_2^2 < \epsilon, \tag{3}$$

where N is the number of occupied blocks, ϵ is the approximation error, controlling the sparsity of $\mathbf{c}_i$, and $\|\cdot\|_0$ is the ℓ_0-norm of vector, counting the number of nonzero elements of the input. Note that $\mathbf{S}_i$ can be adaptively determined only using the geometric data so there is no overhead needed to record $\mathbf{S}_i$. Taking $\mathbf{S}_i\mathbf{\Phi}$ as a whole, i.e., $\widehat{\mathbf{\Phi}}_i = \mathbf{S}_i\mathbf{\Phi} \in \mathbb{R}^{n_i\times k^3}$ ($n_i \ll k^3$), we can see that Eq. (3) is consistent with the well-known problem of SR over an overcomplete (or redundant) dictionary, the overcompleteness of which can achieve a more flexible, more stable, more robust, or more compact expression for signals.

We consider two types of basis $\mathbf{\Phi}$: (1) the DCT basis, i.e., the Kronecker product of three 1D DCT of length $k \times k$; (2) the GT basis over all voxels of a block, whose weight matrix is defined by the inverse distance model, i.e.,

$$w_{ij} = \begin{cases} 1/d(\mathbf{v}_i, \mathbf{v}_j), & \text{if } 0 < d(\mathbf{v}_i, \mathbf{v}_j) < d_{max} \\ 0, & \text{otherwise,} \end{cases} \tag{4}$$

where $d(\mathbf{v}_i, \mathbf{v}_j)$ is the Euclidean distance between two voxels.

For simplicity, we employ the widely-used Orthogonal Matching Pursuit (OMP) algorithm [14] to solve (3) in a

[1]Here, we consider only a single color channel. The other two channels can be processed in the same way.

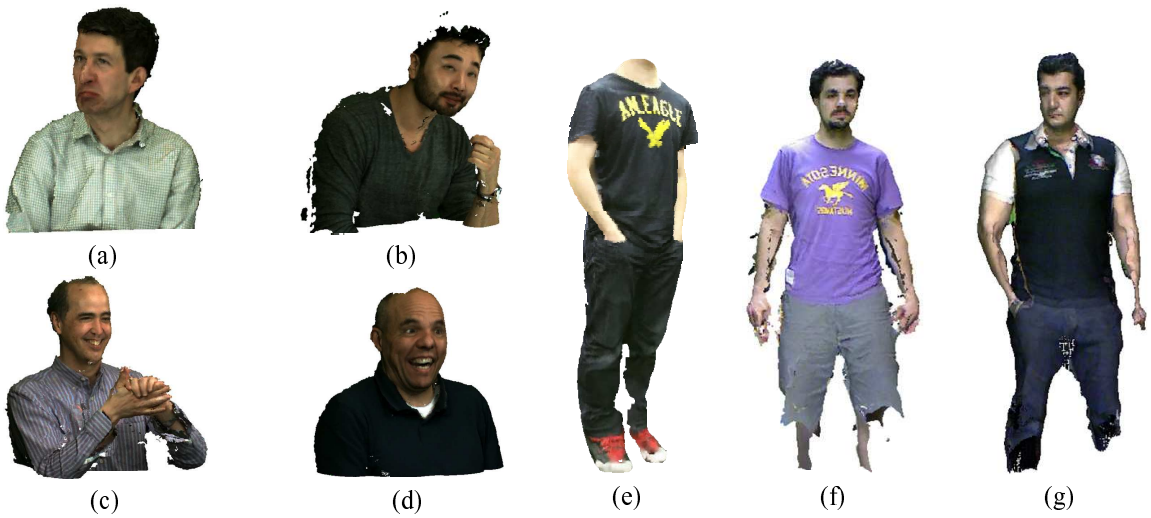

Fig. 1. Some test 3D point clouds rendered in a specific view. (a) *Andrew* (279664, 4072). (b) *David* (330797, 5109). (c) *Phil* (371313, 5732). (d) *Ricardo* (207242, 3053). (e) *Boy* (55489, 2411). (f) *Christos* (155720, 3337). (g) *Dimitris* (131187, 3734). For each model, the two numbers indicate the numbers of occupied voxels and occupied 3D blocks, respectively.

block-by-block manner. Since the M-term approximation behavior (i.e., the error or quality obtained when representing a signal with M nonzero coefficients) of different blocks may be different, one can further improve the sparsity by solving (3) in its Lagrangian form, i.e.,

$$\min_{\{\mathbf{c}_i\}} \frac{1}{2} \sum_{i=1}^{N} \|\mathbf{x}_i - \mathbf{S}_i \mathbf{\Phi} \mathbf{c}_i\|_2^2 + \lambda \sum_{i=1}^{N} \|\mathbf{c}_i\|_0, \tag{5}$$

where $\lambda > 0$ is a penalty parameter, controlling the sparsity of $\mathbf{c}_i$: the larger the value of λ is, the sparser the vector $\mathbf{c}_i$ is. The problem in (5) is equivalent to the summation of multiple independent univariate minimization problems:

$$\sum_{i=1}^{N} \min_{\mathbf{c}_i} \frac{1}{2} \|\mathbf{x}_i - \mathbf{S}_i \mathbf{\Phi} \mathbf{c}_i\|_2^2 + \lambda \|\mathbf{c}_i\|_0, \tag{6}$$

and the subproblems can be separately solved using the iterative hard thresholding algorithm or more advanced proximal methods and alternating direction method of multipliers (ADMM) [15].

4. EXPERIMENTAL RESULTS

We have carried out experiments using frames extracted from sequences of dynamic point cloud data sets: four human upper body frames, i.e., *Andrew*, *David*, *Phil*, and *Ricardo* [16], which have been captured according to [9], and three human full body frames, i.e., *Christos*, *Dimitris* [17], and *Boy*[2]. Figure 1 shows the test data. The seven frames were voxelized by setting L to 9, which yields a $512 \times 512 \times 512$ voxel space[3]. We further partitioned the voxel cube of size $512 \times 512 \times 512$ into $64 \times 64 \times 64$ blocks each of size $8 \times 8 \times 8$.

We compared the proposed algorithm with two recent methods, i.e., GT [4] and Modified SA-DCT [5]. For GT and the proposed algorithm, we tested them under different graph structures corresponding different bases by respectively setting d_{max}^2 to $1, 2, 3$. For GT and Modified SA-DCT, hard thresholding was applied to obtain exactly sparse coefficients[4]. Following existing works in SR, the M-term approximation is adopted to evaluate different algorithms, where the value of M is normalized by the number of occupied voxels and is denoted as p_c, and the reconstruction quality is measured by peak signal-to-noise ratio (PSNR). A better algorithm is one that requires fewer nonzero coefficients at the same reconstruction quality.

Figure 2 shows the experimental results, where the values of PSNR and p_c correspond to the average of three color channels, and we can observe that: (1) our algorithm is always better than the other two algorithms on all seven point clouds, and the superiority becomes more obvious with the reconstruction quality increasing; (2) for *David* and *Ricardo*, our algorithm is comparable to GT at low and medium reconstruction quality, and becomes better but not significantly at higher quality since the textures of these two point clouds are quite smooth, and GT can decorrelate them very well; (3) the performance of our algorithm is only slightly reduced with the increase of d_{max}, indicating that it is more insensitive to the graph structure than GT; our algorithm produces comparable performance under the two types of $\mathbf{\Phi}$; (4) the performance of Modified SA-DCT is much worse than others, and the reason is that too many unoccupied voxels were filled using zeros, leading to many more high frequency transform coeffi-

[2] http://www.kscan3d.com/gallery/

[3] The four human upper body frames provided in the dataset have been voxelized using the real-time high resolution sparse voxelization algorithm [9].

[4] Thresholding yields the best M-term approximation of a signal with respect to an orthonormal basis.

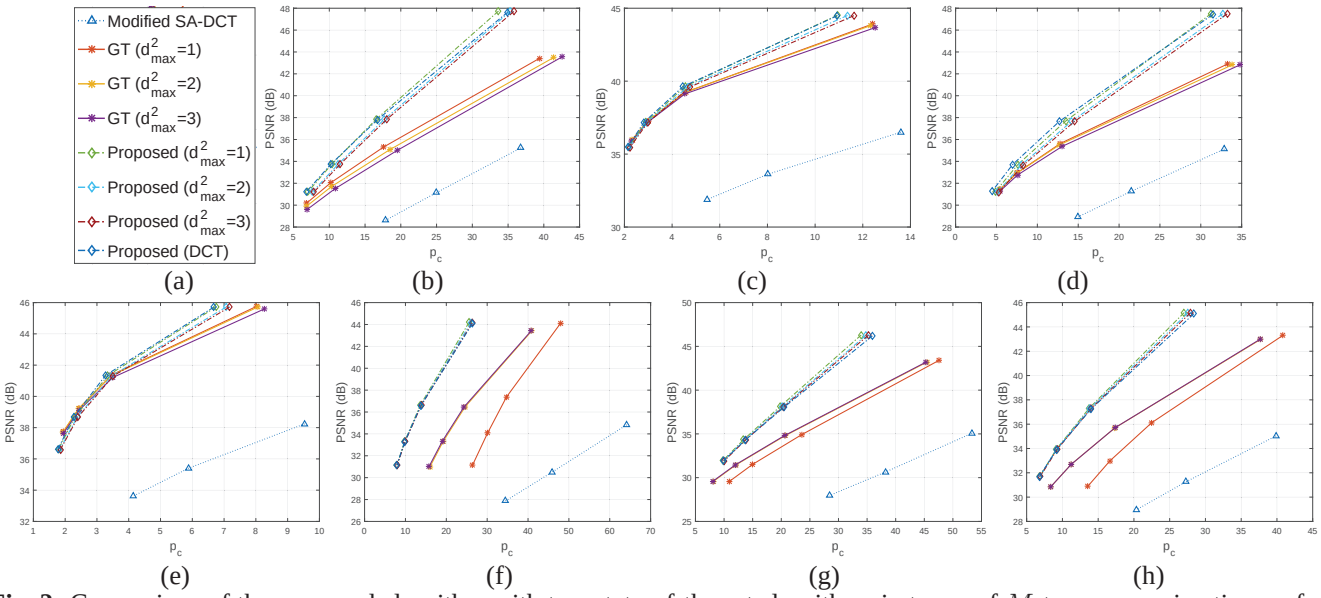

Fig. 2. Comparison of the proposed algorithm with two state-of-the-art algorithms in terms of M-term approximation performance. Note that p_c (%) is computed as the the ratio of the number of nonzero coefficients to the number of occupied voxels. (a) Legend. (b) *Andrew*. (c) *David*. (d) *Phil*. (e) *Ricardo*. (f) *Boy*. (g) *Christos*. (h) *Dimitris*.

cients; the higher compression performance by the Modified SA-DCT based codec compared to the GT-based codec shown in [5] may have benefited from the intra prediction. Alternatively, the Modified SA-DCT can be effective only when used together with a intra prediction.

5. DISCUSSION

In this section, we intuitively explain why the proposed algorithm works using compressive sensing (CS) theory [18], [19], [20].

Consider a sparse or approximately sparse vector $\boldsymbol{\alpha} \in \mathbb{R}^{n\times 1}$. Let $\mathbf{g} = \boldsymbol{\Psi}\boldsymbol{\alpha} \in \mathbb{R}^{m\times 1}$ ($m \ll n$) be the measurement of $\boldsymbol{\alpha}$ via a random measurement (or sensing) matrix $\boldsymbol{\Psi} \in \mathbb{R}^{m\times n}$. CS theory states that if the restricted isometry property (RIP)[5] holds for $\boldsymbol{\Psi}$, then a reconstruction of $\boldsymbol{\alpha}$ can be obtained by solving

$$\widehat{\boldsymbol{\alpha}} \triangleq \arg\min_{\boldsymbol{\alpha}} \|\boldsymbol{\alpha}\|_0 \ \text{ subject to } \ \mathbf{g} = \boldsymbol{\Psi}\boldsymbol{\alpha}. \tag{7}$$

Also, the solution $\widehat{\boldsymbol{\alpha}}$ to (7) obeys

$$\|\widehat{\boldsymbol{\alpha}}-\boldsymbol{\alpha}\|_2 \le C\cdot\|\boldsymbol{\alpha}-\boldsymbol{\alpha}_q\|_1/\sqrt{q} \,\text{ and }\, \|\widehat{\boldsymbol{\alpha}}-\boldsymbol{\alpha}\|_1 \le C\cdot\|\boldsymbol{\alpha}-\boldsymbol{\alpha}_q\|_1 \tag{8}$$

for some constant C, where $\boldsymbol{\alpha}_q$ is the vector $\boldsymbol{\alpha}$ with all but the largest q elements (in magnitude) set to 0. If $\boldsymbol{\alpha}$ is q-sparse (i.e., has at most q nonzero entries $q < m$), then $\widehat{\boldsymbol{\alpha}} = \boldsymbol{\alpha}_q$ and thus the recovery is exact. If $\boldsymbol{\alpha}$ is not strict q-sparse, then (8) asserts that the quality of the recovered signal is as good as if one knew ahead of time the location of the q largest values of $\boldsymbol{\alpha}$ and decided to measure those directly.

[5]We refer readers to [21] for the rigorous definition of RIP. Intuitively speaking, the RIP is to say that all subsets of m columns taken from $\boldsymbol{\Psi}$ are nearly orthogonal.

Remark: According to CS theory, we can conclude that if the RIP holds for $\mathbf{S}_i\boldsymbol{\Phi}$, the optimized sparse coefficients $\mathbf{c}_i$ can represent $\mathbf{y}_i$ very well, and likewise $\mathbf{x}_i$ since $\mathbf{x}_i$ is a subset of $\mathbf{y}_i$. Moreover, we only care about whether $\mathbf{x}_i$ can be well represented instead of $\mathbf{y}_i$, which is different from the goal of CS to recover the unknown original signal, and thus, much sparser coefficients in (3) can be expected, although the RIP is not completely satisfied in practice.

6. CONCLUSION AND FUTURE WORK

We have presented a very effective algorithm to sparsely represent colors on unstructured 3D point clouds. We adopt a virtual adaptive sampling process to bridge elegantly the gap between the advanced SR techniques for regular signals and unstructured 3D point clouds. Moreover, the proposed algorithm can be intuitively explained by compressive sensing theory. Experimental results demonstrate its superiority over state-of-the-art methods. Specifically, the proposed algorithm produces up to 5 dB higher reconstruction quality using the same number of nonzero coefficients or the same reconstruction quality using up to 45% fewer nonzero coefficients.

In the future, we would like to explore the potential of the proposed algorithm along the following directions:
(1) Using dictionary learning techniques to obtain data content adaptive transform matrices $\boldsymbol{\Phi}$;
(2) Integrating the proposed algorithm with other techniques such as quantization, entropy coding, and predictive coding to develop a complete codec for compressing static/dynamic point clouds; and
(3) Denoising the attributes and inpainting the attributes of holes or damaged parts on point clouds (assuming the geometry is first filled using existing methods, e.g., [22]).

7. REFERENCES

[1] 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.

[2] 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. PP, no. 99, pp. 1–1, 2016.

[3] F. Lozes, A. Elmoataz, and O. Lézoray, "Pde-based graph signal processing for 3-d color point clouds: Opportunities for cultural heritage," *IEEE Signal Processing Magazine*, vol. 32, no. 4, pp. 103–111, 2015.

[4] C. Zhang, D. Florêncio, and C. Loop, "Point cloud attribute compression with graph transform," in *Proc. IEEE International Conference on Image Processing (ICIP)*, 2014, pp. 2066–2070.

[5] R. Cohen, D. Tian, and A. Vetro, "Point cloud attribute compression using 3-d intra prediction and shape-adaptive transforms," in *Proc. Data Compression Conference (DCC)*, 2016, pp. 1–10.

[6] D. Thanou, P. A. Chou, and P. Frossard, "Graph-based compression of dynamic 3d point cloud sequences," *IEEE Transactions on Image Processing*, vol. 25, no. 4, pp. 1765–1778, 2016.

[7] B. Dado, T. R. Kol, P. Bauszat, J.-M. Thiery, and E. Eisemann, "Geometry and attribute compression for voxel scenes," *Computer Graphics Forum*, vol. 35, no. 2, pp. 397–407, 2016.

[8] R. Schnabel and R. Klein, "Octree-based point-cloud compression," in *Proc. Eurographics Symposium on Point-Based Graphics*, 2006, pp. 111–120.

[9] C. Loop, C. Zhang, and Z. Zhang, "Real-time high-resolution sparse voxelization with application to image-based modeling," in *Proceedings of the 5th High-Performance Graphics Conference*. ACM, 2013, pp. 73–79.

[10] R. Rubinstein, A. M. Bruckstein, and M. Elad, "Dictionaries for sparse representation modeling," *Proc. of the IEEE*, vol. 98, no. 6, pp. 1045–1057, 2010.

[11] J. Mairal, F. Bach, and J. Ponce, "Sparse modeling for image and vision processing," *Foundations and Trends® in Computer Graphics and Vision*, vol. 8, no. 2-3, pp. 85–283, 2014.

[12] M. Elad, "Sparse and redundant representation modelingłwhat next?" *IEEE Signal Processing Letters*, vol. 19, no. 12, pp. 922–928, 2012.

[13] T. Sikora and B. Makai, "Shape-adaptive dct for generic coding of video," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 5, no. 1, pp. 59–62, 1995.

[14] Y. C. Pati, R. Rezaiifar, and P. Krishnaprasad, "Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition," in *Proc. Asilomar Conference on Signals, Systems and Computers*, 1993, pp. 40–44.

[15] F. Bach, R. Jenatton, J. Mairal, and G. Obozinski, "Optimization with sparsity-inducing penalties," *Foundations and Trends® in Machine Learning*, vol. 4, no. 1, pp. 1–106, 2012.

[16] C. Loop, Q. Cai, S. Orts 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.

[17] A. Doumanoglou, D. S. Alexiadis, D. Zarpalas, and P. Daras, "Toward real-time and efficient compression of human time-varying meshes," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 24, no. 12, pp. 2099–2116, 2014.

[18] E. Candes and J. Romberg, "Sparsity and incoherence in compressive sampling," *Inverse problems*, vol. 23, no. 3, p. 969, 2007.

[19] E. J. Candès and M. B. Wakin, "An introduction to compressive sampling," *IEEE Signal Processing Magazine*, vol. 25, no. 2, pp. 21–30, 2008.

[20] E. J. Candes, J. K. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Communications on pure and applied mathematics*, vol. 59, no. 8, pp. 1207–1223, 2006.

[21] E. J. Candes and T. Tao, "Decoding by linear programming," *IEEE Transactions on Information Theory*, vol. 51, no. 12, pp. 4203–4215, 2005.

[22] Y. Quinsat *et al.*, "Filling holes in digitized point cloud using a morphing-based approach to preserve volume characteristics," *International Journal of Advanced Manufacturing Technology*, vol. 81, no. 1-4, pp. 411–421, 2015.