Multimedia Tools and Applications
https://doi.org/10.1007/s11042-019-7545-0

Variable block-size compressed sensing for depth map coding

Ying Liu[1] · **Joohee Kim**[2]

Received: 11 June 2018 / Revised: 18 March 2019 / Accepted: 21 March 2019 /
Published online: 18 April 2019

Abstract
Compressed sensing (CS) is the theory and practice of sub-Nyquist sampling of sparse signals of interest. Perfect reconstruction is possible with much fewer than the Nyquist required number of data samples. In this work, we consider a variable block-size CS architecture for fast compression of depth maps for three-dimensional video (3DV) applications. While existing CS-based depth map coding methods encode depth maps with equal block size, the proposed algorithm partitions a depth map into smooth and edge blocks of variable sizes via rate-distortion optimized quad-tree decomposition. CS is then performed on edge blocks, and eight-bit encoding is performed on smooth blocks. At the decoder, high quality depth map reconstruction is achieved by minimizing the spatial total-variation. Experimental results show that at a small extra expense of encoder complexity, the proposed variable block-size compressed sensing has enhanced significantly the rate-distortion performance over existing low-complexity CS-based depth map coding algorithms.

Keywords Compressed sensing · Depth map · Quad-tree decomposition · Rate-distortion optimization · Three-dimensional video · Total-variation

1 Introduction

Recent advance in display and camera technologies has enabled three-dimensional video (3DV) applications such as 3D television and stereoscopic cinema. In order to provide the "look-around" effect that audiences expect from a realistic 3D scene, a vast amount of multiview video data needs to be stored or transmitted, leading to the desire of efficient compression techniques. One proposed solution is to encode several selected views of the

✉ Ying Liu
yliu15@scu.edu

Joohee Kim
joohee@ece.iit.edu

[1] Santa Clara University, Santa Clara CA, USA

[2] Illinois Institute of Technology, Chicago IL, USA

same scene captured from different viewpoints along with the corresponding depth (disparity) maps. With texture video sequences and depth map sequences, an arbitrary number of intermediate views can be synthesized at the decoder side using depth image-based rendering (DIBR) techniques [20]. Depth maps, therefore, are considered as an essential coding target for 3DV applications. Typically, depth maps can be well approximated as piecewise smooth signals, with relatively constant depth areas separated by sharp edges where each smooth depth region may correspond to an object at a different depth. Many existing methods have utilized these characteristics for high efficiency depth map coding. For instance, linear functions are constructed to effectively represent smooth areas [17], shape-adaptive wavelet transform is developed for explicit encoding of the locations of major edges [16], and edge adaptive transforms are developed to avoid filtering across edges and to create small coefficient values [19].

While all aforementioned methods rely on encoders of high complexity, new low complexity depth map encoders are recently developed under compressed sensing (CS) framework. CS is an emerging body of work that deals with sub-Nyquist sampling of sparse signals of interest [3–5]. Rather than sampling signals at the Nyquist rate, CS collects only a few (random [4] or deterministic [8]) linear measurements, and the computational burden for successful reconstruction of the original high dimensional signal is shifted to the receiver side. The receiver relies on effective sparse representations of the original signal and appropriate recovery algorithms such as convex optimization [2], linear regression [7, 21], or greedy recovery algorithms [22]. Since depth maps contain piecewise smooth areas with very few texture details, their pixel gradients along horizontal and vertical directions are highly sparse, therefore CS can be considered for fast depth map encoding, while high quality decoding is achievable at the decoder with gradient-sparsity constraint. Such a setup may be of particular interest in problems where low-complexity encoding is required and increased decoder complexity is affordable. A typical example is large wireless sensor networks for 3D surveillance, where power-limited cameras need to be deployed to capture 3D scenes which are sent to a central server or remote viewer for off-line processing.

In existing CS-based depth map coding methods, the depth image is partitioned into blocks of equal size for CS acquisition. Although low-complexity encoding is achieved, the sampling procedure is still highly redundant since there are large smooth or uniform areas of irregular shapes. In this present work, we propose a variable block-size CS architecture which partitions the smooth and edge areas of a depth map into blocks of variable sizes via rate-distortion (RD) optimized quad-tree decomposition. CS is then performed on edge blocks with partial 2D DCT sensing matrix, and eight-bit encoding is performed on smooth blocks for enhanced coding efficiency. At the decoder, high quality depth map reconstruction is achieved by exploiting the sparsity of the pixel domain gradient.

The remainder of this paper is organized as follows. In Section 2, we briefly review the related work on CS-based depth map coding. In Section 3, the proposed variable block-size CS depth map coding architecture is proposed with corresponding reconstruction algorithm. Experimental results and performance analysis are presented in Section 4. Finally, a few conclusions are drawn in Section 5.

2 Related work

An existing CS-based depth map coding algorithm [18] proposed to compress a single depth map via a random subsampling matrix of the Fourier transform basis, and reconstruct the depth map by applying the pixel-domain total-variation (TV) constraint and the Fourier

transform domain sparsity constraint. Interestingly, Duan et al. [6] reveals the fact that when TV constraint is imposed, it is possible to reconstruct the depth image from its partial 2D DCT samples with much higher quality than pure inverse partial 2D DCT. In such scenario, the partial 2D DCT sensing matrix offers much higher coding efficiency than random sensing matrices adopted in general CS community. The drawback of random sensing matrices for compression purposes is that they generate CS measurements of high entropy and lead to low coding efficiency. Meanwhile, the TV constraint at the decoder ensures the smoothness of the depth map and preserves the discontinuities at the edges at the same time. Nevertheless, the algorithm in [18] was designed only for single depth map compression rather than for depth map sequences, and the algorithm in [6] is based on fixed block size CS acquisition, hence redundant sampling occurs for large smooth areas with trivial depth value variations.

On the other hand, graph-based transform (GBT) has also been proposed for CS-based depth map coding [10]. In such scheme, partial Hadamard transform is used as the sensing matrix and GBT is constructed per depth image block as the sparse basis. Although the GBT provides more effective sparse representation for depth maps than other orthogonal basis such as 2D DCT, the construction of block-adaptive GBT increases encoder complexity, and the side information needed to specify the GBT at the decoder heavily increases the required transmission bandwidth. In our preliminary work [15], quad-tree decomposition is utilized to partition a depth map into uniform blocks of variable sizes and small edge blocks of a fixed size. Lossless eight-bit encoding is then applied to each uniform block and only the edge blocks are encoded with CS. Such scheme improves coding efficiency compared to equal block-size CS encoding by avoiding repeated sampling of large uniform areas of irregular shape where all pixels have exactly the same intensity value.

In this paper, an improved variable block-size CS encoder is proposed for depth map coding based on RD optimized quad-tree decomposition. Rather than decomposing a depth map block based on its uniformness as in [15], the encoder computes the encoded bit rate and estimated reconstruction distortion of the original block, as well as the total bit rate and total estimated reconstruction distortion of its four sub-blocks. A Lagrangian functional is then utilized to determine the total cost of the block in both decomposed and non-decomposed cases so as to select the best coding mode. Such RD optimized quad-tree decomposition results in smooth and edge blocks of variable sizes, which provides more flexibility than compressively sampling edge blocks of one fixed size in [15].

3 Proposed methods

3.1 Variable block-size intra-frame CS encoder

We first introduce an intra-frame depth map encoder based on variable block-size CS (VCS). In general, each depth image is virtually partitioned into non-overlapping macro blocks $\mathbf{Z}$ of equal size $n \times n$, which are then encoded individually using the proposed variable block-size CS. As shown in Fig. 1, via an L-level bottom-up RD optimized quad-tree decomposition, each macro block $\mathbf{Z}$ is first decomposed into smooth blocks of size $n_s \times n_s$ and edge blocks of size $n_e \times n_e$, where $n_s, n_e \in \{n \times 2^{1-\ell} \mid \ell = 1, 2, ..., L\}$. To enhance coding efficiency, a smooth block can be approximated as a uniform region represented by a single value, which still preserves the quality of the synthesized views. Meanwhile, accurate representation of edges in depth maps is more important because errors in edge information may lead to significant quality degradation in the synthesized views. Hence, we propose to encode an

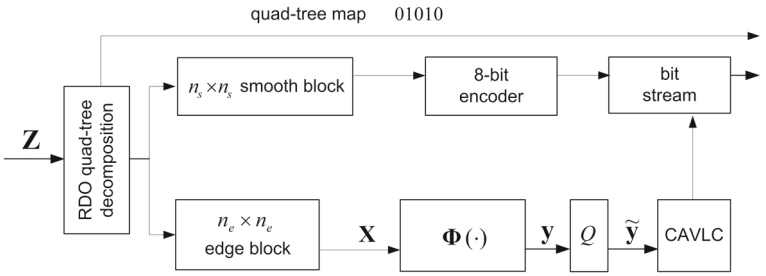


Fig. 1 Intra-frame variable block-size CS encoder

edge block with CS followed by entropy coding, and reconstruct it via edge preserving convex optimization at the decoder side. In particular, if the standard deviation of all the pixel intensity values in one block does not exceed a threshold η, the block is determined as a smooth block and encoded with eight-bit representation which stands for the average intensity value over all pixels in the block. For each edge block $\mathbf{X}$, CS is performed in the form of

$$\mathbf{y} = \mathbf{\Phi}(\mathbf{X}), \tag{1}$$

where $\mathbf{\Phi}(\cdot)$ is the so-called partial 2D DCT sensing operator that generates the top P frequency components of the zig-zag scanned 2D DCT coefficients. Then, the resulting measurement vector $\mathbf{y} \in \mathbb{R}^P$ is processed by a scalar quantizer with certain quantization parameter (QP), and the quantized indices $\widetilde{\mathbf{y}}$ are entropy encoded using context-adaptive variable length coding (CAVLC) as in standard video coding and transmitted to the decoder.

The advantage of such CS depth map encoder lies in the RD optimized quad-tree decomposition. To obtain a global optimal quad-tree decomposition of the depth map, we adopt an L-level bottom-up tree pruning technique. The guiding principle is to parse the initial full tree from bottom (the L^{th} level) to top (the 1^{st} level) and recursively prune leaf nodes (i.e. merge blocks) of the tree according to a decision criterion. The bit rate and distortion calculation of a leaf node $\mathbf{X}_\ell$ on the ℓ^{th} level is shown in Fig. 2. If $\mathbf{X}_\ell$ is determined as a smooth block, its bit rate is 8 and its distortion is approximated as 0. Otherwise, its bit rate is the number of bits after CS acquisition, quantization and entropy coding, and its distortion is estimated by the sum absolute difference (SAD) between the original block and the block recovered from de-quantization $\mathbf{Q}^{-1}$ and inverse partial 2D DCT CS operation $\mathbf{\Phi}^{-1}(\cdot)$.

The tree pruning criterion for the bottom (L^{th}) level leaf nodes $\mathbf{X}_L^j$, $j = 1, 2, 3, 4$ that share the same parent node is shown in Fig. 3. The bit rates and distortions of the children nodes are first computed as R_L^j and D_L^j, $j = 1, 2, 3, 4$ with intra-frame rate and distortion computation (IRDC) as depicted in Fig. 2. Afterwards, the bit rate and distortion

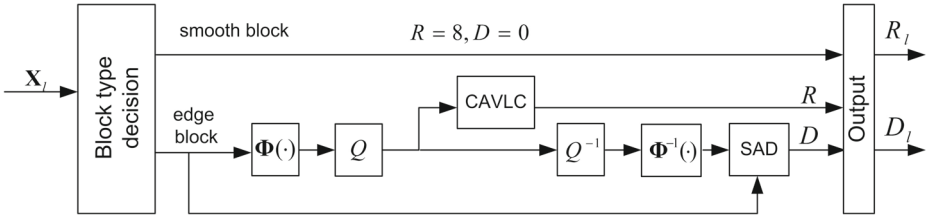


Fig. 2 I frame rate and distortion computation (IRDC) of an ℓ^{th} level leaf node $\mathbf{X}_\ell \in \mathbb{R}^{n_\ell \times n_\ell}$

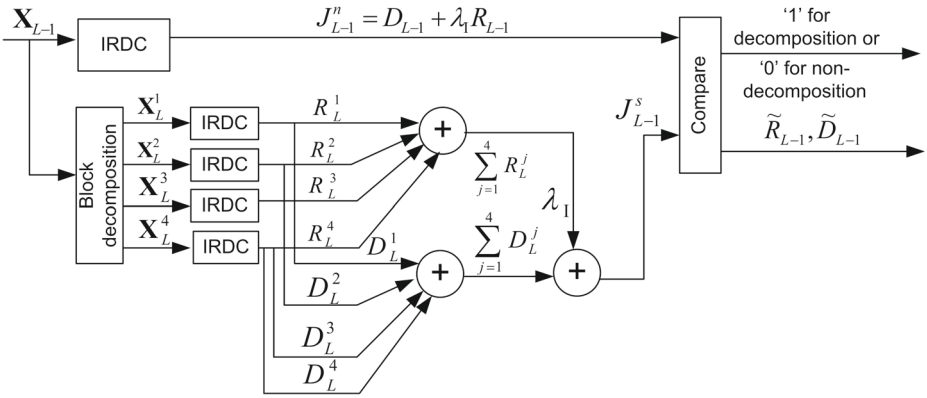


Fig. 3 Pruning criterion for I frame L^{th} level leaf nodes $\mathbf{X}^j_L \in \mathbb{R}^{n_L \times n_L}$, $j = 1, 2, 3, 4$

of the parent node $\mathbf{X}_{L-1}$ are computed as R_{L-1} and D_{L-1}. The RD optimized quad-tree decomposition criterion then refers to the minimization of the cost function $\mathrm{J} = \mathrm{D} + \lambda\mathrm{R}$. For computational simplicity, we use a fixed Lagrange multiplier λ_{I} for I frame, and compute the non-decomposed cost as $\mathrm{J}^n_{L-1} = \mathrm{D}_{L-1} + \lambda_{\mathrm{I}}\mathrm{R}_{L-1}$, and the decomposed cost as $\mathrm{J}^s_{L-1} = \sum_{j=1}^{4} \mathrm{D}^j_L + \lambda_{\mathrm{I}} \sum_{j=1}^{4} \mathrm{R}^j_L$. If $\mathrm{J}^n_{L-1} > \mathrm{J}^s_{L-1}$, then a "1" is transmitted to indicate $\mathbf{X}^j_L$, $j = 1, 2, 3, 4$ are not merged, and $\widetilde{R}_{L-1} = \sum_{j=1}^{4} \mathrm{R}^j_L$, $\widetilde{D}_{L-1} = \sum_{j=1}^{4} \mathrm{D}^j_L$ are stored as the optimal rate and distortion at the parent node $\mathbf{X}_{L-1}$. Otherwise, a "0" is transmitted to indicate the children nodes are merged into one parent node, and $\widetilde{R}_{L-1} = \mathrm{R}_{L-1}$, $\widetilde{D}_{L-1} = \mathrm{D}_{L-1}$ are stored as the optimal rate and distortion at $\mathbf{X}_{L-1}$.

For $1 \le \ell \le L-2$, the tree pruning criterion for $(\ell+1)^{\text{th}}$ level node $\mathbf{X}^j_{\ell+1}$, $j = 1, 2, 3, 4$ is depicted in Fig. 4. First, the bit rate and distortion of the parent node $\mathbf{X}_\ell$ are computed as R_ℓ and D_ℓ via IRDC, then the optimal bit rates and distortions of its four children blocks which were stored previously as the results of higher level node pruning are summed as $\sum_{j=1}^{4} \widetilde{\mathrm{R}}^j_{\ell+1}$ and $\sum_{j=1}^{4} \widetilde{\mathrm{D}}^j_{\ell+1}$. Next, the non-decompopsed cost is computed as $\mathrm{J}^n_\ell = \mathrm{D}_\ell + \lambda_{\mathrm{I}}\mathrm{R}_\ell$,

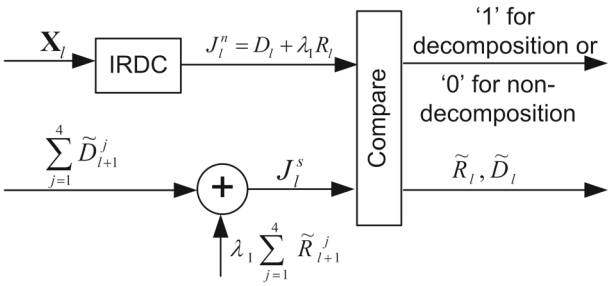


Fig. 4 Pruning criterion for I frame $(\ell+1)^{\text{th}}$ level leaf nodes $\mathbf{X}^j_{\ell+1} \in \mathbb{R}^{n_{\ell+1} \times n_{\ell+1}}$, $1 \le \ell \le L-2$, $j = 1, 2, 3, 4$

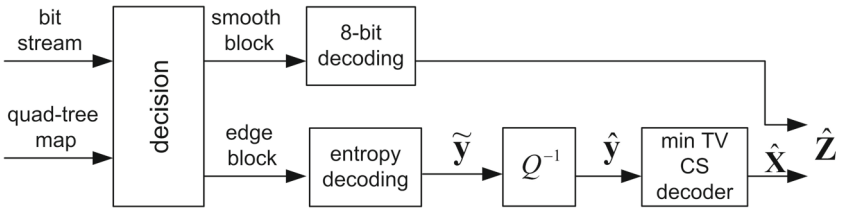


Fig. 5 Intra-frame TV minimization decoder

and the decomposed cost is computed as $\mathrm{J}_\ell^s = \sum_{j=1}^{4} \widetilde{\mathrm{D}}_{\ell+1}^j + \lambda_\mathrm{I} \sum_{j=1}^{4} \widetilde{\mathrm{R}}_{\ell+1}^j$. If $\mathrm{J}_\ell^n > \mathrm{J}_\ell^s$, $\mathbf{X}_{\ell+1}^j$, $j = 1, 2, 3, 4$ are not merged, otherwise they are merged into one parent node $\mathbf{X}_\ell$. Again, the resultant optimal bit rate $\widetilde{\mathrm{R}}_\ell$ and distortion $\widetilde{\mathrm{D}}_\ell$ are stored at node $\mathbf{X}_\ell$ for later use in the ℓ^{th} level node pruning.

Such RD optimized block decomposition algorithm described in Figs. 2 – 4 is performed recursively from the L^{th} level all the way up to the first level of the tree. The resulting bit stream is transmitted as the "quad-tree map" to inform the decoder of the tree pruning structure for successful decoding.

3.2 Total-variation minimization reconstruction

At the decoder, the reconstruction of each macro block is performed independently. As described in Fig. 5, the decoder first reads the bit stream along with the binary quad-tree map to identify smooth and edge blocks. For smooth blocks, a simple eight-bit decoding is carried out. For edge blocks, the decoder performs entropy decoding to obtain the quantized partial 2D DCT coefficients (or CS measurements) $\widetilde{\mathbf{y}}$. The elements of $\widetilde{\mathbf{y}}$ are then de-quantized to form vector $\widehat{\mathbf{y}}$. Since depth map blocks containing edges have sparse spatial gradients, they can be reconstructed via pixel-domain 2D TV[1] minimization in the form of

$$\widehat{\mathbf{X}} = \arg\min_{\mathbf{X}} \mathrm{TV}_{2\mathrm{D}}(\mathbf{X}) \tag{2}$$

$$\text{subject} \quad \text{to} \quad ||\widehat{\mathbf{y}} - \mathbf{\Phi}(\mathbf{X})||_{\ell_2} \leq \epsilon. \tag{3}$$

The reconstructed smooth blocks and edge blocks are re-grouped thereafter to form the decoded macro block $\widehat{\mathbf{Z}}$.

3.3 Extension of inter-frame variable block-size CS coding

So far, we have carried out intra-frame only depth map coding. To exploit temporal correlation among successive frames, we now extend the proposed VCS to inter-frame depth map coding. Since we are targeting at a low-complexity 3DV encoder, some powerful coding tools in standard video coding such as the high complexity motion estimation and delay-sensitive I-B-P coding structure are not considered in this context. Instead, simple frame difference is taken as the residual signal for fast inter-frame coding. At the encoder, the sequence of depth maps is divided into groups of pictures (GOP) with I-P-P-P structure. For each GOP, the intra-frame RD optimized quad-tree decomposition is applied to every macro

[1]The mathematical expression of pixel-domain 2D total-variation is defined as in [13, 14].

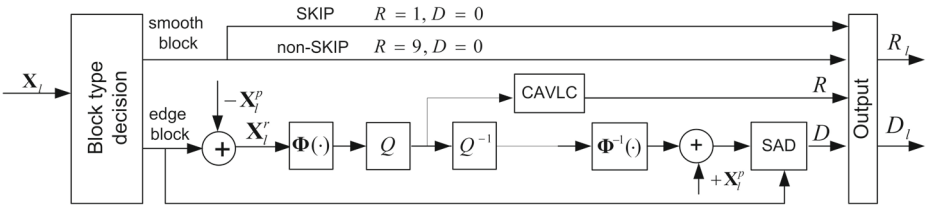


Fig. 6 P frame rate and distortion computation (PRDC) of an ℓ^{th} level leaf node $\mathbf{X}_\ell \in \mathbb{R}^{n_\ell \times n_\ell}$

block of the I frame. For the subsequent P frames, a modified RD optimized quad-tree decomposition is performed.

As shown in Fig. 6, to compute the bit rate and distortion of a P frame ℓ^{th} level leaf node $\mathbf{X}_\ell$, it is first classified as a smooth block or an edge block, and its co-located block in the previous frame recovered from inverse partial 2D DCT CS operation $\mathbf{\Phi}^{-1}(\cdot)$ is extracted as $\mathbf{X}_\ell^p$. A smooth block is further classified as a SKIP block if its mean intensity value is the same as that of $\mathbf{X}_\ell^p$, otherwise it is classified as a non-SKIP block. A SKIP block is not encoded, and the encoder only transmits one bit to indicate the SKIP mode. For a non-SKIP block, the bit rate is $R = 9$, including the one-bit indicator and eight-bit representation of its mean intensity value. For an edge block $\mathbf{X}_\ell$, it is first predicted by $\mathbf{X}_\ell^p$ and the residual block $\mathbf{X}_\ell^r = \mathbf{X}_\ell - \mathbf{X}_\ell^p$ is encoded with the CS operator $\mathbf{\Phi}(\cdot)$, quantized, entropy encoded, and the resulting number of bits is the bit rate. To calculate the distortion, de-quantization is performed followed by inverse CS operation, adding back the reference block $\mathbf{X}_\ell^p$ and SAD computation. The pruning criterion for the P frame leaf nodes on all levels of the tree is similar to the I frame leaf node decomposition shown in Figs. 3 and 4, with all the IRDC modules replaced by the P frame rate and distortion computation (PRDC) module depicted in Fig. 6.

To generalize, the inter-frame encoding procedure is shown in Fig. 7. Denote the k^{th} P frame after the I frame $\mathbf{F}_t$ as $\mathbf{F}_{t+k}$, then a macro block $\mathbf{Z}_{t+k} \in \mathbb{R}^{n \times n}$ in $\mathbf{F}_{t+k}$ can be considered as the input of the RD optimized quad-tree decomposition algorithm. $\mathbf{Z}_{t+k}$ is first partitioned into smooth blocks and edge blocks of variable sizes, while each smooth block is either skipped or encoded with eight bits, for an edge block $\mathbf{X}_{t+k}$, its residual $\mathbf{X}_{t+k}^r$ is compressed using partial 2D DCT sensing matrix to form a residual measurement vector $\mathbf{y}_{t+k}^r$. Afterwards, $\mathbf{y}_{t+k}^r$ is quantized, encoded with CAVLC and transmitted.

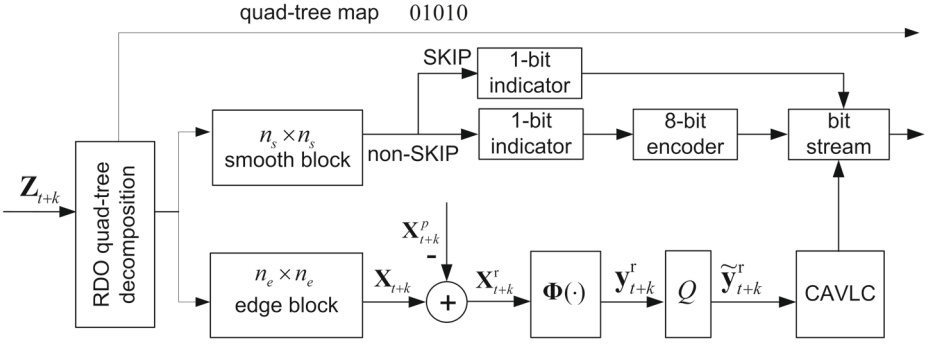


Fig. 7 Inter-frame variable block-size CS encoder

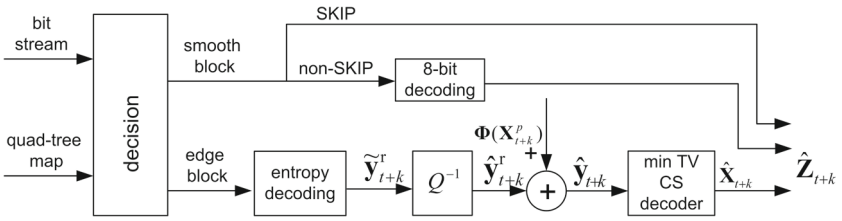


Fig. 8 Inter-frame variable block-size CS decoder

For reconstruction at the decoder, details are shown in Fig. 8. Every SKIP block is decoded as a uniform block with all pixel intensities the same as the mean of its co-located block $\mathbf{X}^p_{t+k}$ in the reference frame, and for every non-SKIP smooth block, eight-bit decoding is performed. For edge block $\mathbf{X}_{t+k}$, the de-quantized residual CS measurement vector $\widehat{\mathbf{y}}^r_{t+k}$ is added to the CS encoded reference signal to approximate the CS measurement vector

$$\widehat{\mathbf{y}}_{t+k} = \mathbf{\Phi}(\mathbf{X}^p_{t+k}) + \widehat{\mathbf{y}}^r_{t+k}. \tag{4}$$

Since the CS acquisition of $\mathbf{X}_{t+k}$ can be formulated as

$$\begin{aligned} \mathbf{y}_{t+k} &= \mathbf{\Phi}(\mathbf{X}_{t+k}) \\ &= \mathbf{\Phi}(\mathbf{X}^p_{t+k} + \mathbf{X}^r_{t+k}) \\ &= \mathbf{\Phi}(\mathbf{X}^p_{t+k}) + \mathbf{y}^r_{t+k} \\ &= \mathbf{\Phi}(\mathbf{X}^p_{t+k}) + \widehat{\mathbf{y}}^r_{t+k} + \mathbf{n}^r_{t+k} \\ &= \widehat{\mathbf{y}}_{t+k} + \mathbf{n}^r_{t+k}, \end{aligned} \tag{5}$$

where $\mathbf{n}^r_{t+k}$ is the noise due to the quantization of $\mathbf{y}^r_{t+k}$, the pixel block $\mathbf{X}_{t+k}$ can be reconstructed via pixel-domain TV minimization in the form of

$$\widehat{\mathbf{X}}_{t+k} = \arg\min_{\mathbf{X}} \mathrm{TV}_{2\mathrm{D}}(\mathbf{X}) \tag{6}$$

$$\text{subject} \quad \text{to} \quad ||\widehat{\mathbf{y}}_{t+k} - \mathbf{\Phi}(\mathbf{X})||_{\ell_2} \leq \epsilon. \tag{7}$$

Finally, the reconstructed smooth blocks and edge blocks are re-grouped to form the decoded macro block $\widehat{\mathbf{Z}}_{t+k}$.

4 Experimental results and performance analysis

In this section, we evaluate the performance of the proposed VCS depth map coding system by comparing the perceptual quality of the decoded depth maps, the RD performance of the synthesized view and the computational complexities of VCS to other depth map coding schemes with low-complexity encoders.

4.1 Experiment set-up

Two test video sequences, *Kendo* and *Balloons*, with a resolution of 1024×768 pixels are used. For both sequences, 40 frames of the depth maps of view 1 and view 3 are compressed using the proposed VCS encoder, and the reconstructed depth maps at the decoder are used to synthesize the texture video sequence of view 2 with the View Synthesis Reference Software (VSRS) [23].

In our experimental studies, the macro block size is $n = 128$, and a five-level ($L = 5$) RD optimized quad-tree decomposition is implemented, resulting in smooth and edge blocks of size $n_\ell \times n_\ell, n_\ell \in \{8, 16, 32, 64, 128\}$. The standard deviation threshold used for smooth and edge block classification is determined empirically as $\eta = 2$, and the Lagrangian multiplier for I frame and P frame are $\lambda_{\mathrm{I}} = 1$ and $\lambda_{\mathrm{P}} = 3.5$, respectively. The CS ratio for each edge block is fixed at $\frac{P_\ell}{n_\ell^2} = 0.375$, where P_ℓ is the number of CS samples for the ℓ^{th} level edge block. Four quantization parameters QP $= \{24, 28, 32, 36\}$ are used to generate different bit rates. To reconstruct edge blocks from partial 2D DCT CS measurements, TVAL3 software [9, 12] is employed to solve the TV minimization problems in (2) and (6).

The proposed VCS depth map coding system is examined with GOP size $T = 20$ with I-P-P-P coding structure. For comparison studies, we include three existing low-complexity depth map encoders: *i*) quad-tree partitioned inter-frame CS encoder without RDO (QCS [15]); *ii*) equal block-size inter-frame CS encoder with TV minimization decoding (ECS [6]); and *iii*) intra-frame CS encoder based on partial Hadamard sensing matrix with GBT sparsifying basis (Intra GBT [10]). For fair comparison, CAVLC is used as the entropy coding scheme for all four encoders. To solve the ℓ_1 minimization problem in the GBT-based algorithm, ℓ_1-magic software [1] is utilized.

4.2 Perceptual quality and RD performance

Figure 9 shows the different reconstructions of the 8th frame of *Kendo* view 1 depth map sequence produced by the proposed VCS (Fig. 9c), the QCS (Fig. 9d), the ECS (Fig. 9e),

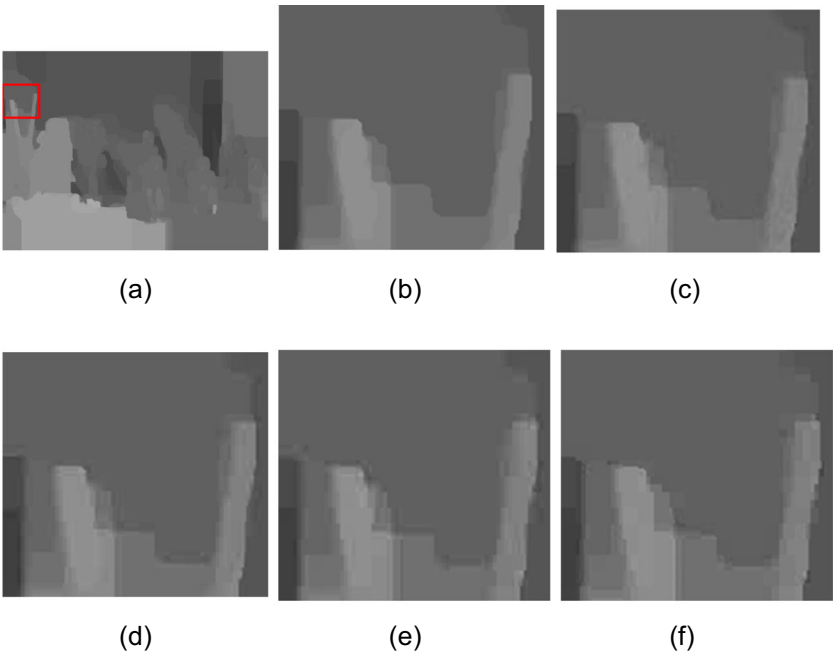


Fig. 9 Different reconstructions of the 8th frame of *Kendo* depth map 1: **a** original *Kendo* depth image; **b** magnified view of the marked area; **c** marked area coded with VCS, $T = 20$ at 0.0787 bpp and 48.8784 dB of PSNR; **d** with QCS, $T = 20$ at 0.073 bpp and 46.4534 dB of PSNR [15]; **e** with ECS, $T = 20$ at 0.1475 bpp and 44.4933 dB of PSNR [6]; and **f** with Intra GBT at 0.233 bpp and 44.1326 dB of PSNR [10]

and the Intra GBT (Fig. 9f) systems. It can be observed that the ECS system as well as the Intra GBT system suffer noticeable performance loss around the edge areas, while the proposed VCS system and the QCS system demonstrate considerable reconstruction quality improvement. Although the perceptual quality difference between VCS and QCS decoding is minor due to the pdf formatting of the present article, VCS actually provides 2.43 dB higher PSNR than QCS decoding at similar bit rates as explained in the figure captions.

Figure 10 shows the rate-distortion characteristics of the *Kendo* sequence. The bit-rate indicates the average bits per pixel (bpp) of the encoded depth map sequences from view 1 and view 3, and the peak signal-to-noise ratio (PSNR) of the luminance component of synthesized view 2 is computed between the rendered view using compressed depth sequences and using the ground-truth depth sequences. Evidently, for a fixed PSNR, VCS with TV minimization (VCS, TV min) outperforms QCS with gains as much as 0.02 bpp. In addition, both VCS and QCS outperform significantly the ECS and GBT coding schemes. To justify the superiority of TV minimization decoding over the direct inverse partial 2D DCT, we also provide the RD curve of VCS encoder with inverse partial 2D DCT decoder (VCS, i2D-DCT). It can be observed that with the same bit rates, VCS with TV minimization decoding improves the PSNR by 0.5 to 1.6 dB compared to inverse partial 2D DCT decoding.

The same perceptual quality evaluation and rate-distortion performance study are performed in Figs. 11 and 12 for the *Balloons* sequence. Similar conclusions can be drawn that our proposed VCS encoder with TV minimization decoding outperforms the other low-complexity depth map encoders.

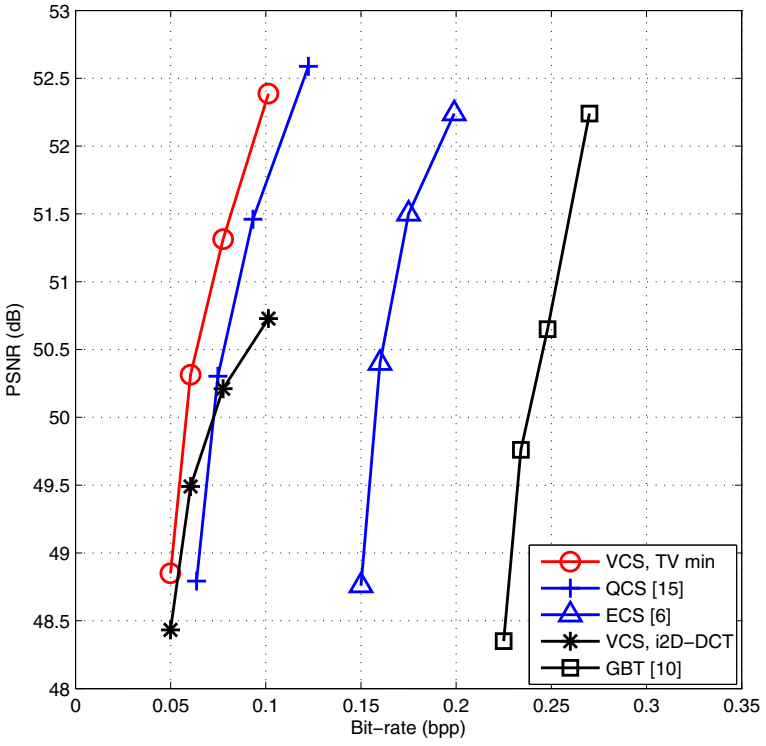


Fig. 10 Rate-distortion studies on the synthesized view 2 of the *Kendo* sequence

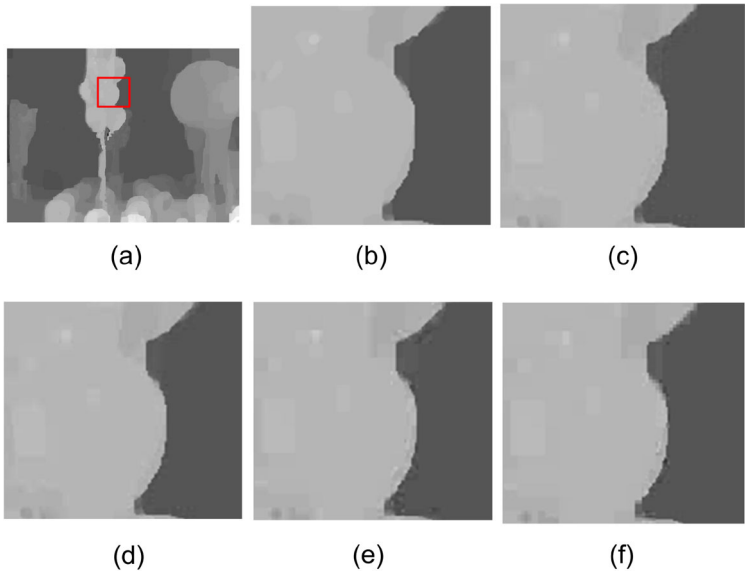


Fig. 11 Different reconstructions of the 2^{nd} frame of *Balloons* depth map 1: **a** original *Balloons* depth image; **b** magnified view of the marked area; **c** marked area coded with VCS, $T = 20$ at 0.1037 bpp and 46.21 dB of PSNR; **d** with QCS, $T = 20$ at 0.1160 bpp and 46.0555 dB of PSNR [15]; **e** with ECS, T=20 at 0.1495 bpp and 41.8531 dB of PSNR [6]; and **f** with Intra GBT at 0.2725 bpp and 41.2059 dB of PSNR [10]

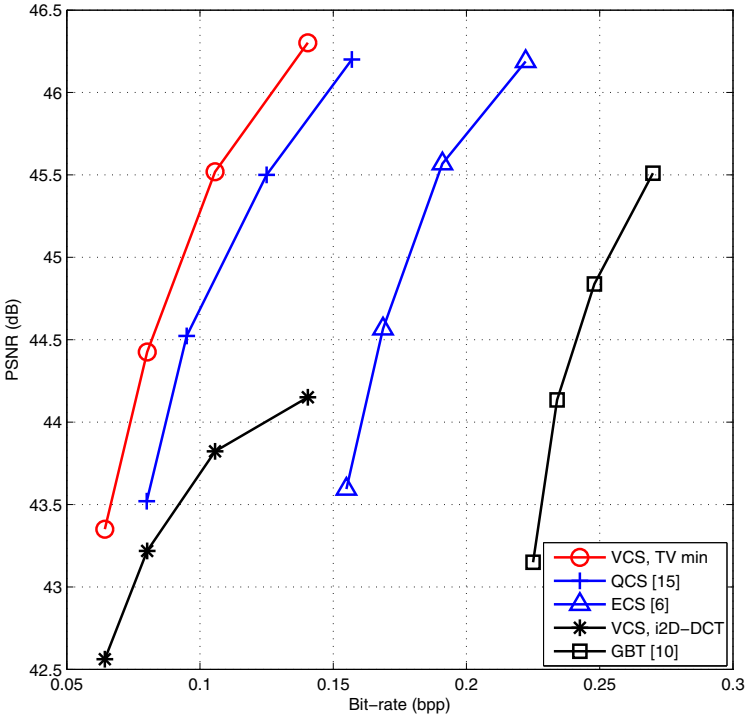


Fig. 12 Rate-distortion studies on the synthesized view 2 of the *Balloons* sequence

Table 1 Encoder complexity per $n \times n$ macro block

Encoder	Block type	Block size	Complexity (or upper bound)
VCS	edge	$\frac{n}{2^{\ell-1}} \times \frac{n}{2^{\ell-1}}, 1 \le \ell \le L$	$\le (r+1)n^3 \sum_{\ell=1}^{L} 2^{1-\ell}$
QCS	edge	$\frac{n}{2^{L-1}} \times \frac{n}{2^{L-1}}$	$\le (r+1)n^3 2^{1-L}$
ECS	arbitrary	$\frac{n}{2^{L-1}} \times \frac{n}{2^{L-1}}$	$(r+1)n^3 2^{1-L}$

4.3 Encoder complexity

The computational cost of the proposed bottom-up RD optimized VCS encoder is analyzed as follows and summarized in Table 1. We consider the encoder complexity for an $n \times n$ macro block. For an ℓ^{th} level sub-block of size $\frac{n}{2^{\ell-1}} \times \frac{n}{2^{\ell-1}}$, the partial 2D DCT CS operation with compressive sampling ratio $r = \frac{P_\ell}{(\frac{n}{2^{\ell-1}})^2}$ takes $(r+1)\left(\frac{n}{2^{\ell-1}}\right)^3$ multiplications. Since there are $4^{\ell-1}$ sub-blocks of size $\frac{n}{2^{\ell-1}} \times \frac{n}{2^{\ell-1}}$ on the ℓ^{th} level, the total complexity for the partial 2D DCT CS operation on the ℓ^{th} level is $(r+1)n^3 2^{1-\ell}$ per $n \times n$ macro block. For QCS in [15], partial 2D DCT is applied to only edge blocks on the L^{th} level, thus the encoder complexity per $n \times n$ macro block is upper bounded by $(r+1)n^3 2^{1-L}$. In the proposed VCS encoder in this work, the complexity of the encoder is upper bounded by $(r+1)n^3 \sum_{\ell=1}^{L} 2^{1-\ell}$ multiplications per $n \times n$ macro block due to the RD optimization on all L levels.

Although the encoder complexity of VCS is increased compared to QCS and ECS, a lot of blocks on each level of the tree are classified as smooth blocks in the RD optimization process for which the CS operations are avoided. As shown in Table 2, the actual number of edge blocks per frame averaged over 40 frames of the *Balloons* depth map 1 sequence of each level of the tree is less than the total number of blocks on that tree level, especially when the block size becomes smaller, leading to a practical encoder complexity much lower than the upper bound in Table 1. Hence, VCS is indeed a low-complexity encoder compared to the H.264 standard video coding which requires sophisticated motion estimation.

4.4 Decoder complexity

The complexity of decoder for the proposed VCS encoder lies in the TV minimization algorithm for edge blocks. As elaborated in [11], the TV minimization algorithm relies on two nested iterations to minimize an augmented Lagrangian cost function. If the number of inner and outer iterations are K and N, respectively, then the total cost for TV minimization is $\mathcal{O}\left(N(K+1)(r+1)(\frac{n}{2^{\ell-1}})^3\right)$ per $\frac{n}{2^{\ell-1}} \times \frac{n}{2^{\ell-1}}$ edge block . Therefore, the decoder complexity can be summarized in Table 3. Although VCS with TV minimization (VCS, TV min) is

Table 2 Average number of edge blocks per frame of *Balloons* depth map 1 sequence

tree level ℓ	1	2	3	4	5
block size	128×128	64×64	32×32	16×16	8×8
number of edge blocks	41	139	414	1118	2682
total number of blocks	48	192	768	3072	12288

Table 3 Decoder complexity per $n \times n$ macro block

Decoder	Block type	Block size	Complexity
VCS, TV min	edge	$\frac{n}{2^{\ell-1}} \times \frac{n}{2^{\ell-1}}, 1 \leq \ell \leq L$	$\mathcal{O}\left(N(K+1)(r+1)n^3\right)$
QCS, TV min	edge	$\frac{n}{2^{L-1}} \times \frac{n}{2^{L-1}},$	$\mathcal{O}\left(N(K+1)(r+1)n^3 2^{1-L}\right)$
ECS, TV min	arbitrary	$\frac{n}{2^{L-1}} \times \frac{n}{2^{L-1}}$	$\mathcal{O}\left(N(K+1)(r+1)n^3 2^{1-L}\right)$
VCS, i2D-DCT	edge	$\frac{n}{2^{\ell-1}} \times \frac{n}{2^{\ell-1}}, 1 \leq \ell \leq L$	$\mathcal{O}\left((r+1)n^3\right)$

more complex than VCS with inverse 2D DCT (VCS, i2D-DCT) due to iterative decoding, it provides much better reconstruction quality as shown in the RD curves in Figs. 10 and 12.

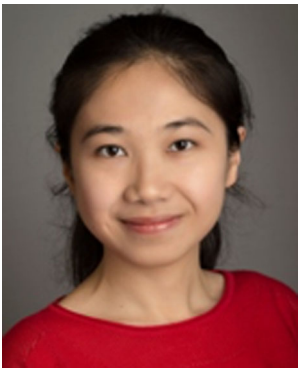
5 Conclusions

We proposed a low-complexity variable block-size CS architecture for depth map coding. In particular, a five-level bottom-up quad-tree decomposition is developed using recursive rate-distortion optimization to partition the depth map into smooth and edge blocks of variable sizes. While each smooth block is encoded using efficient eight-bit approximation that results to negligible distortion, the edge blocks are encoded with partial 2D DCT CS operator. At the decoder, total-variation minimization which enforces sparse gradient constraint is utilized for CS encoded edge block reconstruction. Experimental results demonstrate that the proposed VCS depth map coding greatly enhances the RD performance of the non-RD optimized quad-tree partitioned CS coding, the equal block-size CS coding, and the intra-frame GBT based CS coding at a small extra expense of encoder complexity. Moreover, TV minimization decoding provides better depth map reconstruction quality than direct inverse 2D DCT decoding. In terms of future work, motion information can be exploited at the decoder to further enhance the signal sparsity, hence leading to better reconstruction quality.

References

1. Candès E, Romberg J $\ell 1$-magic: recovery of sparse signals via convex programming, www.acm.caltech.edu/l1magic/downloads/l1magic.pdf
2. Candès E, Romberg J, Tao T (2006) Stable signal recovery from incomplete and inaccurate measurements. Commun Pure and Appl Math 59(8):1207–1223
3. Candès E, Tao T (2006) Near optimal signal recovery from random projections: universal encoding strategies? IEEE Trans Inf Theory 52(12):5406–5425
4. Candès E, Wakin MB (2008) An introduction to compressive sampling. IEEE Signal Processing Mag 25(2):21–30
5. Donoho DL (2006) Compressed sensing. IEEE Trans Inf Theory 52(4):1289–1306
6. Duan J, Zhang L, Liu Y, Pan R, Sun Y (2011) An improved video coding scheme for depth map sequences based on compressed sensing. In: Proceedings of the international conference on multimedia technology (ICMT), Hangzhou, China, pp 3401–3404
7. Efron B, Hastie T, Johnstone I, Tibshirani R (2004) Least angle regression. Ann Statist 32:407–451
8. Gao K, Batalama SN, Pados DA, Suter BW (2011) Compressive sampling with generalized polygons. IEEE Trans Signal Process 59(10):4759–4766
9. Jiang H, Li C, Haimi-Cohen R, Wilford P, Zhang Y (2012) Scalable video coding using compressive sensing. Bell Labs Tech J 16:149–169
10. Lee S, Ortega A (2012) Adaptive compressed sensing for depthmap compression using graph-based transform. In: Proceedings of IEEE international conference on image process (ICIP), Orlando, FL, pp 929–932

11. Li C (2009) An efficient algorithm for total variation regularization with applications to the single pixel camera and compressive sensing, Rice University
12. Li C, Jiang H, Wilford P, Zhang Y (2011) Video coding using compressive sensing for wireless communications. In: Proceedings of IEEE wireless communications & networking conference (WCNC) Cancun, Mexico, pp 2077–2082
13. Liu Y, Pados DA (2013) Decoding of framewise compressed-sensed video via interframe total variation minimization. SPIE J Electron Imaging, Special Issue on Compressive Sensing for Imaging, vol 22, no 2
14. Liu Y, Pados DA (2013) Rate-adaptive compressive video acquisition with sliding-window total-variation-minimization reconstruction. In: Proceedings of SPIE, compressive sensing conference, SPIE defense, security, and sensing, Baltimore, MD, vol 8717
15. Liu Y, Vijayanagar KR, Kim J (2014) Quad-tree partitioned compressed sensing for depth map coding. In: IEEE international conference on acoustics, speech and signal process (ICASSP), Florence, Italy, pp 870–874
16. Maitre M, Do MN (2010) Depth and depth-color coding using shape-adaptive wavelets. J Vis Commun Image R 21(5-6):513–522
17. Morvan Y, de With PHN, Farin D (2006) Platelet-based coding of depth maps for the transmission of multiview images. In: Proceedings of SPIE stereoscopic displays and virtual reality systems XIII, vol 6055, p 60550K
18. Sarkis M, Diepold K (2009) Depth map compression via compressed sensing. In: Proceedings of IEEE international conference on image process (ICIP), Cairo, Eygpt, p 737C740
19. Shen G, Kim W-S, Narang SK, Ortega A, Lee J, Wey H (2010) Edge-adaptive transforms for efficient depth-map codin. In: Proceedings of picture coding symposium (PCS), Nagoya, Japan, pp 566–569
20. Smolic A, Müller K, Dix K, Merkle P, Kauff P, Wiegand T (2008) Intermediate view interpolation based on multiview video plus depth for advanced 3D video systems. In: Proceedings of the IEEE international conference on image process (ICIP), San Diego, CA, pp 2448–2451
21. Tibshirani R (1996) Regression shrinkage and selection via the lasso. J Roy Stat Soc Ser B 58(1):267–288
22. Tropp J, Gilbert A (2007) Signal recovery from random measurements via orthogonal matching pursuit. IEEE Trans Inf Theory 53(12):4655–4666
23. View synthesis reference software (VSRS 3.5), In: Tech. Rep. ISO/IEC JTC1/SC29/WG11 (2010)



Ying Liu received the B.S. degree in Communications Engineering from Beijing University of Posts and Telecommunications, Beijing, China, in 2006, the M.S. and Ph.D. degrees in Electrical Engineering from The State University of New York at Buffalo, Buffalo, NY, USA, in 2008 and 2012, respectively. She currently is an Assistant Professor in the Department of Computer Engineering at Santa Clara University, Santa Clara, CA, USA. Her general areas of expertise are computer vision, machine learning, and signal processing.

Joohee Kim received the B.S. and M.S. degrees in Electrical and Electronic Engineering from Yonsei University, Seoul, Korea in 1991 and 1993, respectively. She received the Ph.D. degree in Electrical and Computer Engineering from the Georgia Institute of Technology, Atlanta, GA, in 2003.

From 1993 to 1997, she was with Korea Telecom Research Laboratory, Seoul, Korea as a Research Engineer. She joined Samsung Advanced Institute of Technology, Suwon-si, Korea in 2003 as a Senior Research Engineer and developed various video coding algorithms. From 2005 to 2008, she was an Assistant Professor in the Department of Information and Communication Engineering at Inha University in South Korea. She joined the faculty of the Illinois Institute of Technology (IIT), Chicago, IL, in 2009 and is currently an Associate Professor of the Department of Electrical and Computer Engineering. She is the Director of Multimedia Communications Laboratory at IIT and has been actively involved in research projects funded by US Federal Agencies and Korean Government. Her current research interests include image and video signal processing, multimedia communication, multimedia systems, 3D video representation and transmission, real-time 3D reconstruction for augmented teleoperation, computer vision and machine learning.