IET Image Processing

Research Article

IET Journals
The Institution of Engineering and Technology

View synthesis method for 3D video coding based on temporal and inter view correlation

ISSN 1751-9659
Received on 17th November 2017
Revised 03rd May 2018
Accepted on 16th July 2018
doi: 10.1049/iet-ipr.2018.5390
www.ietdl.org

Long Vuong Tung[1] ✉, Minh Le Dinh[1], Xiem HoangVan[1], Trieu Duong Dinh[1], Tien Huu Vu[2], Ha Thanh Le[1]

[1]*VNU University of Engineering and Technology, Vietnam National University, 144 Xuan Thuy, Cau Giay, Hanoi, Vietnam*

[2]*Posts and Telecommunications Institute of Technology, Km10, Nguyen Trai, Ha Dong, Hanoi, Vietnam*

✉ *E-mail: longvt94@gmail.com*

Abstract: Recently, in three-dimensional (3D) television, the temporal correlation between consecutive frames of the intermediate view is used together with the inter-view correlation to improve the quality of the synthesised view. However, most temporal methods are based on the motion vector fields (MVFs) calculated by the optical flow or block-based motion estimation which has very high computational complexity. To alleviate this issue, the authors propose a temporal-disparity-based view synthesis (TDVS) method, which uses the MVFs extracted from the bitstreams of side views and motion warping technique to create the temporal correlation between views in the intermediate position. Then a motion compensation technique is used to create a temporal-based view. Finally, the temporal-based view is fused with a disparity-based view which is generated by a traditional depth image-based rendering technique to create the final synthesised view. The fusion of these views is performed based on the side information which is determined and encoded at the sender-side of the 3D video system using a dynamic programming algorithm and rate-distortion optimisation scheme. Experimental results show that the proposed method can achieve the synthesised view with appreciable improvements in comparison with the view synthesis reference software 1D fast (VSRS-1D Fast) for several test sequences.

1 Introduction

Three-dimensional television (3DTV) and free viewpoint video (FVV) are known as two main 3D video applications which have been extensively studied in many fields from video acquisition to display technologies [1]. While the 3DTV provides perceptional depth using tailored displays with or without special glasses, the FVV allows users to interactively navigate their viewpoints. However, 3DTV and FVV generally require a huge number of views to roam around a scene, and due to the limitation in hardware and bandwidth resources, the acquisition and transmission of such a huge number of views are not possible. To solve these problems, the multiple-view video plus depth (MVD) method has been introduced for effective multi-view data representation [2]. In the MVD method, only a few views are captured, coded and transmitted; and then, at the receiver, intermediate views between the existing viewpoints are synthesised so that the burdens for encoding and transmitting colour videos of full viewpoints can be significantly reduced.

Generally, synthesised views are generated from real views by depth image-based rendering (DIBR) techniques [3]. A typical DIBR scenario consists of three main steps: disparity-based warping, view merging, and hole filling. However, due to the lack of original information, the DIBR-based view synthesis is still an ill-posed problem which may create noise or unreliable synthesised data. More specifically, areas that have the same depth and uniform textures are usually represented without distortions, while foreground object edges and more complex textures have a high distortion. Besides, in the case of view merging in DIBR, the boundaries tend to be blurred because of blending colour between background and foreground. Additionally, after view merging, there are remaining holes present in the virtual view. The small holes can be handled based on interpolation or extrapolation techniques. However, simple texture interpolation or extrapolation is insufficient for larger holes, especially holes at the highly textured background. To improve the quality of the synthesised view, several techniques have been integrated into the DIBR mechanism. In [4], Yang *et al.* introduced a reliability reasoning scheme for disparity-based warping which assesses the reliability of each pixel value in the synthesised view. Then the quality of the synthesised view is improved by withdrawing the unreliable pixels from the view. Lee *et al.* [5] proposed a background contour region replacement method to clean background noises in the warped views to improve the quality of the synthesised view. In [6], Muller *et al.* proposed a prioritised multi-layer projection scheme to reduce boundary artefacts. In [7], Criminisi *et al.* proposed a hole filling algorithm based on inpainting techniques [8, 9]. Both the texture and structure from neighboured regions are simultaneously propagated to fill the holes. Ismael *et al.* [10] proposed an extension to the Criminisis algorithm by including the depth information to guide the propagation process. Although giving the better performance than interpolation and extrapolation techniques, the inpainting process results in much higher complexity.

As presented, most of the mentioned view synthesis methods utilise only the inter-view correlations between views to create the synthesised view.

In fact, there is also the temporal correlation between frames inside the synthesised view which can significantly improve the quality of the synthesised view. For instance, in [11–14], the authors compute the motion vector fields (MVFs) of the reference views and warp vectors into the synthesised view. The warped motion vectors are then used to exploit the temporal correlation between frames. More specifically, Purica *et al.* [11] and Chen *et al.* [12] use warped MVFs to retrieve information about dis-occluded regions from other frames. In [13], the motion compensation (MC) is performed with sub-pixel precision using warped MVFs to obtain temporal predictions which are blended together with the DIBR. The view synthesis method proposed in [14] also use warped MVFs to create temporal predictions for the intermediate view, but an additional frame per group of pictures (GOP) in the intermediate view is required to compress and send as the reference frame for the MC. In [15], Minh *et al.* directly estimate forward MVF between frames of intermediate view and use bi-directional motion estimation scheme to convert and refine MVF between the previous and past frame to MVFs from previous and past to current frame. Then, the bi-directional motion compensate is performed to create the temporal prediction frame. However, because of high precise MVFs requirement, these methods need to use the optical flow or block-based motion estimation to deliver MVFs which may not be appropriate for

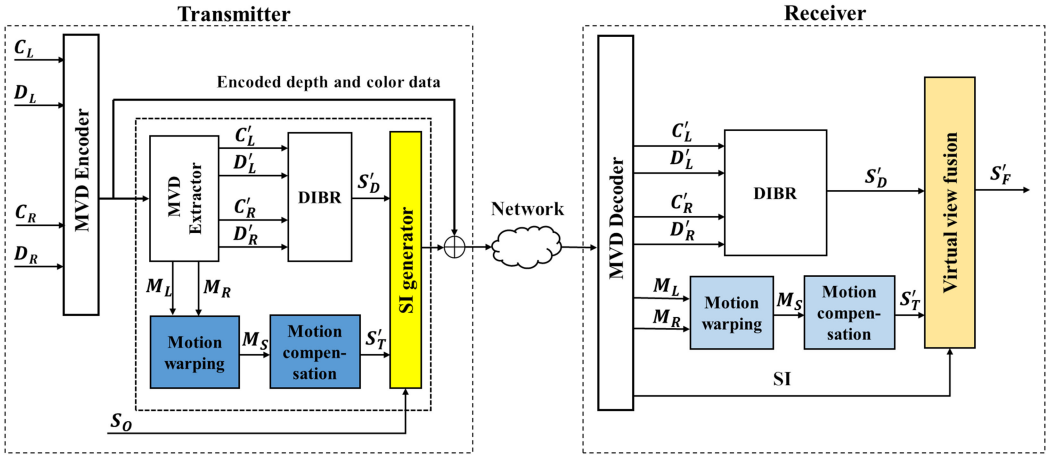

Fig. 1 *Proposed view synthesis framework*

many emerging applications constrained by the computation resources. For example, in [11, 13, 14], the optical flow algorithm in [16] is used to deliver MVFs.

To address these problems, we propose in this study, a novel view synthesis method which exploits not only the inter-view and temporal correlation but also existing MVFs in bitstreams which are calculated at the encoder side to improve the quality of the synthesised view. Note that, in video compression [17], the motion vector is key to reduce redundancy in video data. At the encoder, the MVFs are estimated and fed to an MC process to prediction the next frame. Then, the differences between the prediction frame and real frame along with the MVFs are coded and added to the bitstreams. At the decoder, the received MVFs are used to reconstruct the frame. In the proposed method, two intermediate views, called temporal-based and disparity-based views, are generated and then fused to create the final synthesised view. The temporal-based view is generated using motion warping and MC techniques where MVFs are immediately extracted from the bitstreams of reference views received from the encoder side while the disparity-based view is generated using a traditional DIBR technique. The fusion of these views to create the final synthesised view is performed based on the side information (SI) which is determined and encoded at the sender side of the 3D video system using a dynamic programming algorithm and rate-distortion optimisation (RDO) scheme. Experimental results show that the proposed method provides better quality compared with another conventional disparity-based view synthesis.

The remainder of this paper is organised as follows. Section 2 describes the proposed view synthesis method. Then, Section 3 presents the test conditions and discusses the experimental results. Finally, Section 4 gives the conclusions.

2 Proposed view synthesis method

Fig. 1 illustrates a block diagram of our proposed view synthesis method. The notations in this figure are explained throughout this section while the novel techniques associated with the proposed temporal and inter-view correlation-based view synthesis are described in the next subsections. The proposed view synthesis framework operates as follows:

At the transmitter: A pair of two views (C_{L}, D_{L}) and (C_{R}, D_{R}) denoted for the colour, depth video of left and right views, respectively, are encoded using the MVD encoder and then transmitted to the receiver. Next, the encoded colours (C'_{L}, C'_{R}) and depths (D'_{L}, D'_{R}) along with their MVFs (M_{L}, M_{R}), respectively, are extracted by the MVD extractor. Then, the extracted colours and depths are used to generate the disparity-based view, S'_{D}, using the traditional DIBR algorithm. The temporal-based view, S'_{T}, is created by using the MC algorithm described in the next subsection. Finally, in order to fuse the temporal-based view and the disparity-based view appropriately, side information is generated based on the support of S_{O}, which can be the original view or the synthesised view rendered from uncoded colour and depth at the virtual position.

At the receiver: The decoded colours, depths and motion information, (C'_{r}, D'_{r}, M_{r}), with r in {L, R}, respectively, are exploited to create the disparity-based view S'_{D}, using the traditional DIBR method and the temporal-based view, S'_{T}, using MC technique. Then, the SI received from the encoder is utilised to fuse these two intermediate views to achieve the final synthesised view, S'_{F}, in the virtual view fusion module. More details on the view synthesis fusion are presented in the following subsections.

2.1 Inter view correlation-based virtual view creation

The disparity synthesised views are generated from the colour and depth of the left view (C'_{L}, D'_{L}) and those of the right view (C'_{R}, D'_{R}), respectively, by using the traditional DIBR technique. A typical DIBR scenario consists of three steps: 3D warping, view merging, and hole filling. In this scenario, 3D warping is used to project the pixels of the decoded views C'_{L} and C'_{R} to the target synthesised virtual view, S'_{D}, using depth images D'_{L} and D'_{R}, respectively. Due to the occluded region between decoded views, 3D warping can expose areas called holes. View merging combines all the warped views into one image, resulting in the reduction of holes. The remaining holes in the synthesised view are then handled by the hole filling algorithms, which are generally based on interpolation techniques.

2.2 Temporal correlation-based virtual view creation

The block-based MVFs [18] of the left and right views are extracted from the bitstreams and converted to pixel-based MVFs by assigning the motion vector of blocks to their pixels. Then, these pixel-based MVFs denoted by M_{L} and M_{R} are projected to the virtual view. Fig. 2 shows the relation between the positions of a real point projection in different views and at two different time instants $t-1$ and t.

Let C_{r}^{t-1}, C_{r}^{t}, S^{t-1}, and S^{t} denote the colour frame of the reference views (r = L, R) and synthesised view at time $t-1$ and t, respectively. Let $k^t = (x, y)$ be a point in C_{r}^{t} associated with the motion vector $M_{\mathrm{r}}(k^t)$ which can be estimated during the encoding process of the reference view. The motion vector $M_{\mathrm{r}}(k^t)$ points to a corresponding point $k^{t-1} = k^t + M_{\mathrm{r}}(k^t)$ in C_{r}^{t-1}. Let $M_{\mathrm{S}}(k^t + d^t(k^t))$ be the motion vector of the projection of point k^t onto S^t where $d^t(k^t)$ is the disparity value of the reference view at time t. $M_{\mathrm{S}}(k^t + d^t(k^t))$ can be computed as follows:

$$\begin{aligned} M_{\mathrm{S}}(k^t + d^t(k^t)) &= [k^{t-1} + d^{t-1}(k^{t-1})] - [k^t + d^t(k^t)] \\ &= [k^t + M_{\mathrm{r}}(k^t) + d^{t-1}(k^t + M_{\mathrm{r}}(k^t))] - [k^t + d^t(k^t)] \\ &= M_{\mathrm{r}}(k^t) + d^{t-1}[k^t + M_{\mathrm{r}}(k^t)] - d^t(k^t)\,. \end{aligned}$$

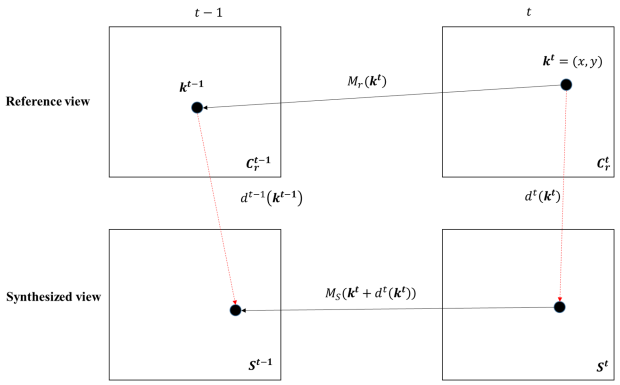

Fig. 2 *Motion warping*

(1)

The disparity value can be calculated from the depth image as follows [19]:

$$d^t(k^t) = f \times L \times \left[\frac{D_r^{'t}(k^t)}{255} \times \left(\frac{1}{Z_{\min}} - \frac{1}{Z_{\max}}\right) + \frac{1}{Z_{\max}}\right], \quad (2)$$

where $D_r^{'t}(k^t)$ denotes a decoded depth value at position k^t of reference view r, f is the focal length, L is the baseline width (horizontal interval between two view points), and $Z_{\min}$ and $Z_{\max}$ are the nearest and farthest object distances of the scene, respectively.

The above calculation only regards the projection of the motion vectors of one reference view to the synthesised view. With multiple reference views, it is possible to have more than one motion vector projected from reference views at one position $k^t + d^t(k^t)$ at the synthesised view. In this case, the projected motion vector from the closest real view is chosen.

When the MVFs of the synthesised view are calculated, the MC will be performed to create a motion compensated frame at the synthesised view denoted as S'_T as seen in Fig. 1. Note that the MC can only be done in the regions where the motion information is available.

2.3 Virtual view fusion

This step combines pixel values from S'_D and S'_T to get an improved synthesised view. For this purpose, we propose in this study a novel *encoder-driven virtual view fusion algorithm* which creates SI at the encoder and embeds this data into the encoded bitstreams to help better fusing the pixel values from S'_D and S'_T at the decoder.

2.3.1 Encoder side-information generation: Since S'_D, S'_T and S_O are available at the encoder, the synthesis algorithm simply selects pixel values from either S'_D or S'_T which is closer to the reference data S_O. Let A_{2D} be the 2D distortion analysis map which indicates the difference between the square errors of the reference data, S_O to the synthesised data S'_D and S'_T as follows:

$$A_{2D}(k) = \begin{cases} 0, & S'_T(k) = 0, \\ [S'_D(k) - S_O(k)]^2 - [S'_T(k) - S_O(k)]^2, & \text{otherwise}, \end{cases} \quad (3)$$

where $k = (x, y)$ is a point that scans every pixel position of the whole image plane and $S'_T(k) = 0$ means that the motion vector at position (x, y) is not available.

Fig. 3 shows an example of the distortion analysis map. As both intra and inter modes are used in the encoding process, only the samples in the inter-coded regions that are associated with the motion vectors can be exploited to create the temporal synthesised view. In the figure, the pixels associated with the intra-coded regions are illustrated as the blank cells.

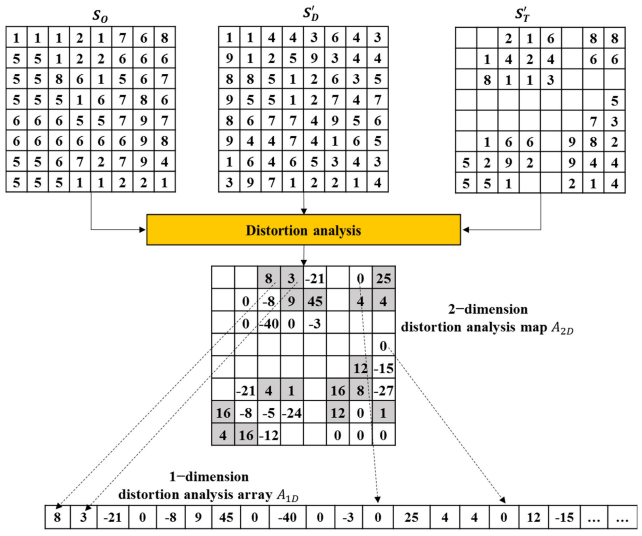

Fig. 3 *Distortion analysis illustration*

It can be seen in (3) that the pixel value $S'_T(k)$ is closer to reference data $S_O(k)$ than the pixel value $S'_D(k)$ if $A_{2D}(k) > 0$. In this case, the synthesised pixel from the temporal view synthesis will be considered as reliable and the location k will be compressed and sent to the decoder. The view synthesis algorithm in the decoder creates the final synthesised view by selecting the synthesised view data from $S'_T(k)$; otherwise, $S'_D(k)$ will be selected.

To efficiently compress the reliable location information, we propose to create a 1D distortion analysis array A_{1D} from the mentioned 2D distortion analysis map A_{2D} and exploit a dynamic programming algorithm to locate non-overlapping continuous sub-arrays. These sub-arrays contain reliable positions in A_{1D}, where a continuous sub-array of A_{1D}, represented by (h, t) includes all elements from $A_{1D}[h]$ to $A_{1D}[t]$ of A_{1D}. Specifically, sub-array (h, t) denotes $\{A_{1D}[h], A_{1D}[h+1], A_{1D}[h+2], \ldots, A_{1D}[t-1], A_{1D}[t]\}$. Then, a sub-array, ith, can be represented by a pair (a_i, b_i) where a_i and b_i are the first and last positions of the sub-array, respectively. These pairs (a_i, b_i) can be entropy coded and signalled to the decoder to present the MC regions instead of coding individually reliable positions. For clarification, the algorithm detail is explained as follows:

- First, the connected components labelling algorithm [20] is applied to A_{2D} to create a list of connected components whose corresponding pixel values in $S'_T > 0$.
- Second, each connected component is scanned with the raster pattern pixel-by-pixel and the scanned pixel values are appended consecutively to A_{1D}. As seen in Fig. 3, four connected components are visible in A_{2D}, and the pixel values of these components are scanned to construct the 1D array A_{1D}.
- Third, for the detection of reliable positions and the guarantee of maximal distortion reduction, we propose a dynamic

programming algorithm that finds N non-overlapping continuous sub-arrays of A_{1D}, $\{(a_i, b_i)|1 \le i \le N\}$, such that the total sum of these sub-arrays, $\sum_{i=1}^{N}\sum_{j=a_i}^{b_i} A_{1D}(j)$, is the largest. We define the optimal-value function $f(n, j, p)$ as the largest sum obtained by selecting n sub-arrays from the first j elements of A_{1D} and $p = \{0, 1\}$, where $p = 1$ indicates that the jth element belongs to nth sub-array and $p = 0$ otherwise. The recursions are defined as follows:

$$f(n, j, 0) = \max\{f(n, j-1, 1), f(n, j-1, 0)\} \tag{4}$$

and

$$f(n, j, 1) = \max\begin{Bmatrix} f(n, j-1, 1), \\ f(n-1, j-1, 1), \\ f(n-1, j-1, 0) \end{Bmatrix} + A_{1D}(j), \tag{5}$$

where

$$\begin{aligned} &\text{initial case:} && f(0, j, p) = 0. \\ &\text{constraints:} && j > n \quad \text{for} \quad f(n, j, 0). \\ &&& j \ge n \quad \text{for} \quad f(n, j, 1). \end{aligned} \tag{6}$$

If the jth element is not included in the nth sub-array, the maximum sum by selecting n sub-arrays from the first j elements is equivalent to that of the first $j-1$ elements. It yields (4) to update $f(n, j, 0)$.

If the jth element is included in the nth sub-array, two situations must be considered. First, if n sub-arrays are selected from the first $j-1$ elements, the $(j-1)$th element must be included in the nth sub-array to ensure the continuation of the nth sub-array. Therefore, $f(n, j, 1) = f(n, j-1, 1) + A_{1D}(j)$. Second, if $n-1$ sub-arrays are selected from the first $j-1$ elements, the jth element must be the first element of the nth sub-array; hence, $f(n, j, 1) = \max\{f(n-1, j-1, 1), f(n-1, j-1, 0)\} + A_{1D}(j)$. The combination of the two situations yields (5).

Fig. 4 shows an example of the sub-array selection and the operation of the dynamic programming algorithm. The initial case and the constraints of the algorithm are highlighted.

- Finally, the largest total sum of N sub-arrays is $\max\{f(N, |A_{1D}|, 0), f(N, |A_{1D}|, 1)\}$, where $|A_{1D}|$ is the total number of elements of A_{1D}.

2.3.2 Rate distortion optimisation for n sub-array: As can be seen from the algorithm, the larger number of sub-arrays used as SI, the smaller distortion yielded in the synthesised view. However, the larger number of sub-arrays used, the larger number of bits needed to present sub-arrays. To find an optimal number of sub-arrays, an RDO mechanism can be applied. In this case, the cost function that balances the rate and distortion based on the Lagrangian optimisation technique [21] can be defined as follows:

$$J(\lambda, N) = D(N) + \lambda \times R(N), \tag{7}$$

where λ is the Lagrange multiplier, N is the number of sub-arrays fed to the proposed dynamic programming algorithm to solve the N maximal sums problem on array A_{1D}, and $R(N)$ is the number of bits needed to present N sub-arrays $\{(a_i, b_i)|1 \le i \le N\}$. The distortion of the fused synthesis view, $D(N)$, is computed as

$$D(N) = D_c - D_T(N), \tag{8}$$

where $D_c = \sum_{i=1}^{H}\sum_{j=1}^{W}[S'_D(i, j) - S_O(i, j)]^2$ is the distortion of the disparity-based synthesised view against S_O and $D_T(N) = \sum_{i=1}^{N}\sum_{j=a_i}^{b_i} A_{1D}(j)$ is the distortion reduction when temporally fusing the synthesised view using n sub-array A_{1D}.

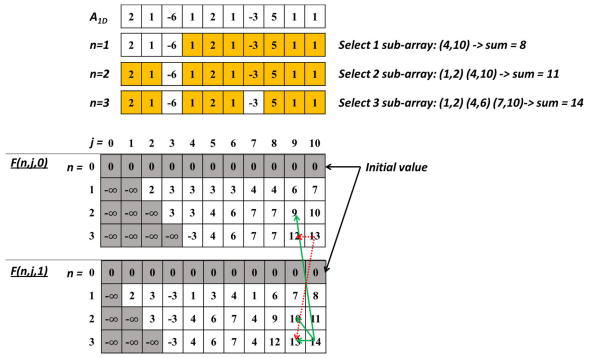

Fig. 4 *Example of the proposed dynamic programming algorithm*

It can be seen that $D_T(N+1) - D_T(N) \le (D_T(N)/N)$ and $(D_T(N+1)/(N+1)) \le (D_T(N)/N)$. Therefore, $D(N)$ decreases when N increases and has a lower bound. Furthermore, because the rate $R(N)$ linearly increases when N increases, $J(\lambda, N)$ is a convex function. The optimal values of λ and N, denoted by $\bar{\lambda}$ and $\bar{N}$, respectively, can be found by solving the following:

$$(\bar{\lambda}, \bar{N}) = \arg\min_{\{\lambda, N\}} J(\lambda, N). \tag{9}$$

It can be seen in (7) and (9) that λ is generally considered to have a great influence on the Lagrange RDO cost function. Thus, in order to achieve the optimal RD performance, it is important to choose a reasonable $\bar{\lambda}$. Its value is usually determined empirically [21–23] as follows:

$$\bar{\lambda} = 0.85 \times 2^{(QP-12)/3}, \tag{10}$$

where (QP) is the quantisation parameter. Therefore, in this work, we test $\bar{\lambda}$ around the value above and it is empirically set as

$$\bar{\lambda} = 1 \times 2^{(QP-12)/3}. \tag{11}$$

As $J(\lambda, N)$ is a convex function, it has only one minimum value. Therefore, $\bar{N}$ is initially set to a number that is small enough, and it is increased by 1 until $J(\bar{\lambda}, \bar{N})$ reaches the minimum and starts to increase. This searching procedure is performed during the execution of the proposed dynamic programming algorithm.

2.3.3 Fusion of the synthesised view at the decoder: At the decoder, the N optimised sub-arrays (a_i, b_i) extracted and decoded from the bitstreams are used as the fusion decision map for the reconstruction of the final synthesised view as follows:

$$S'_F(x, y) = \begin{cases} S'_T(x, y), & (x, y) \in \omega_i, \\ S'_D(x, y), & (x, y) \notin \omega_i, \end{cases} \tag{12}$$

where $\omega_i = \{(x, y)|(x, y)$ is the reverse mapping to the 2D- map at every position k with $k \in \cup_{i=1}^{N}[a_i, b_i]\}$.

3 Results and discussion

To evaluate the effectiveness of the proposed method, we utilise the test model designed for 3D-high efficiency video coding (HEVC) [24]. The video test sequences [25] are Balloons, Kendo, Newspaper and PoznanHall2. Table 1 shows more details on the setting parameters for these test sequences. We encode the left and right views as shown in Table 1 of these sequences separately with QP values of 25, 30, 35, and 40 using HEVC reference software [26]. The intra period is 24, the length of the GOP is set to 8, and the GOP structure is IPP..PPP where frames only reference their previous frame. The traditional view synthesis method based on the DIBR technique is performed by view synthesis reference software 1D fast (VSRS-1D Fast) rendering used in 3D-HEVC standardisation [24] which became an anchor to several new rendering techniques. Furthermore, the original sequences in the

intermediate view are used as references for our proposed method to generate SI in these experiments.

3.1 Virtual view quality assessment

To evaluate the visual quality of virtual views, the synthesis results are compared against the original intermediate sequences through two quality metrics: the peak-signal-to-noise ratio (PSNR) and the structural similarity index (SSIM) [27]. Fig. 5 shows the PSNR variation of the synthesised view over time with the reference and the proposed view synthesis methods while Table 2 shows the PSNR [dB] results for the reference VSRS-1D Fast method, the PSNR improvements [dB] of the proposed TDVS method and the simple average-based fusion (ABF) method for each tested sequence and QP. As shown in Fig. 5, the proposed method outperforms the reference view synthesis VSRS-1D Fast for all tested sequences at all frames. Note in the plots that because there is no motion information in I-Frames at frame numbers: 0, 24, 48 and 72, the PSNR of the proposed method is equal to that of VSRS-1D Fast at those I-Frames. As shown in the summary in Table 2, it is incompetent to combine pixel values from disparity and temporal prediction without SI. The synthesised become worse when using simple ABF. In contrast, the proposed method outperforms the reference view synthesis VSRS-1D Fast method for all tested sequences at all tested QP values with the average gains of 0.33 dB. As the target of SI is to minimise the signal error between reference view and synthesis view, we use an additional quality metric: SSIM which is considered to be correlated with the quality perception of the human visual system to guarantee that our proposed synthesised view is also better in human visual perspective. As shown in the summary in Table 3, the results present that our method still outperforms the reference method for the proposed (0.9475) and reference (0.9469) method on average over the tested sequences for SSIM.

Table 4 shows the percentage of pixels using values from the temporal-based prediction in our method for each tested sequence and QP. It can be seen that the higher percentage of using the temporal synthesis is obtained for the lower QP scenarios. This occurs because the quality of the temporal synthesis block is high when the quality of the left- and right-decoded views is high.

Fig. 6 shows a rendering example of a frame of sequence Newspaper at QP 25 with three enlarged image regions for visual analysis. Note that the quality of the synthesised colour images is degraded in both the proposed and VSRS-1D Fast methods. The artefacts are especially noticeable at the object boundaries as shown in Figs. 6*c*, *d*, *f*, *g*, *i* and *j*. However, the results of the proposed method in Figs. 6*d*, *g* and *j* have fewer artefacts and geometry distortions than those of VSRS-1D Fast in Figs. 6*c*, *f* and *i*, respectively.

3.2 Rate-distortion (RD) performance evaluation

We evaluate the RD performance of the reference and the proposed method using Bjontegaard delta-PSNR [28] metric. The PSNR is evaluated against the original intermediate views. The rate in the reference method is the sum of the rates needed to code the left and right views with their associated depth videos. The rate in the proposed method is the rate for the reference method added to the rate of the SI. We consider two representation types of SI: fixed length (FL) and simple Huffman code (HC) for the experiments in this study. Specifically, for FL representation, $2 \times [\log_2(H \times W)]$ is the number of bits to present a pair (a_i, b_i). For HC representation, the codes are generated over sequences and QPs.

The overheads from the SI for different QPs are shown in Table 5. The RD curves for the VSRS-1D Fast and the proposed methods are given in Fig. 7 while Table 6 shows the Δ PSNR improvement. In Table 6, a positive Δ PSNR value indicates the coding gain in the intermediary view synthesised by the proposed method. As is shown, the proposed view synthesis method achieves a quality improvement of 0.24 dB in the case of FL representation and 0.28 dB in case of HC representation.

3.3 SI efficiency evaluation

As discussed in Section 2.3, the SI is a set of sub-arrays. Each sub-array is represented by a pair of number (a_i, b_i) where the numbers between a_i and b_i present the reliable positions of the temporal-based prediction that would improve the final virtual view. To evaluate the efficiency of the SI generation algorithm, we calculate the average length of sub-arrays by the equation as follows:

$$\frac{1}{N}\sum_{i=1}^{N}(b_i - a_i + 1), \tag{13}$$

where N is the number of selected sub-arrays.

Table 7 shows the average lengths of sub-arrays for different QPs. As can be seen in the table, the average lengths are longer for higher QP values. The reason is that the RDO algorithm only selects sub-arrays that have sufficient contribution to minimising errors between the reference view and the synthesised view in order to ensure the RD performance. In high-QP setting, because of the low quality of the reference data for the temporal-based prediction, the error signal decreases very small for each pixel selected. Thus, the sub-arrays are protracted to deliver adequate improvement. Consequently, the number of sub-arrays in the high-QP setting is less than the low-QP setting but it carries more reliable positions.

3.4 Additional complexity

Regarding the complexity of the proposed methods, additional computation is required for encoding, decoding and rendering. For the decoding process, we need additional complexity for decoding the SI. In our experiments, we consider two types of representation for the SI: FL and HC, so the additional complexity is negligible. For the rendering process, because the MVFs used to generate the temporal prediction are extracted from conventional encode and decode processes of reference views, the additional time complexity is very low for the rendering. Specifically, rendering includes the conversion of MVFs from the block-based to the pixel-based, disparity-based motion warping, MC and fusion. The complexities of these steps are all $O(H \times W)$, where H and W are the height and width of the frame. The additional computations for the encoding process are approximately equal to the sum of additional computations for rendering and the generation of the SI at the encoder. The complexity of the SI generation is $O(N \times H \times W)$, where N is determined by the RDO scheme.

4 Conclusion

In this study, we presented a view synthesis method that exploits not only the inter-view correlation between views but also the temporal correlation within the view to improve the quality of the synthesised images. The contributions of this study include a novel temporal correlation-based virtual view creation method and an encoder-driven view fusion method. Experimental results showed that the proposed view synthesis method outperforms the traditional DIBR-based view synthesis (VSRS-1D Fast) method, notably in both synthesised view quality improvement and RD performance. However, the transmitted fusion decision map is related to only one synthesised view, the bit-rate will increase as transmission of multiple fusion decision maps, in order to be applied to synthesise arbitrary views, that the bit-rate will increase as transmission of multiple fusion decision maps. This is the weak point of our current work. For future work, the extension approach we focus on is only analysing decoded information, especially MVFs of two reference views, to fuse those virtual view frames.

5 Acknowledgments

This work was supported by the project named Multimedia Application Tools for Intangible Cultural Heritage Conservation and Promotion (DTDL.CN-34/16) funded by Vietnam Ministry of Science and Technology.

Table 1 Selected test video sequences

Sequence name	Left view	Synthesis view	Right view	Resolution
balloons	1	3	5	1024 × 768
kendo	1	3	5	1024 × 768
newspaper	2	4	6	1024 × 768
poznanhall2	5	6	7	1920 × 1088

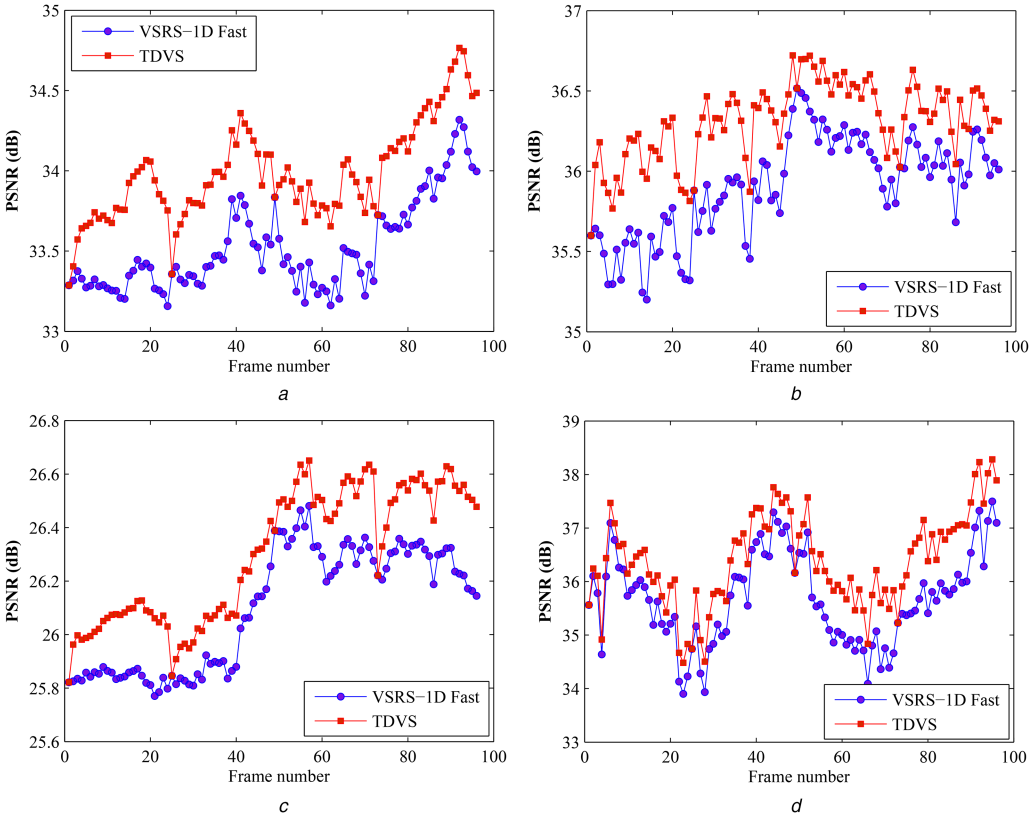

Fig. 5 *PSNR comparison between the traditional and proposed methods at QP 25*

Table 2 PSNR performance with different QPs

Sequence	VSRS-1D fast				ABF versus VSRS-1D Fast				TDVS versus VSRS-1D Fast			
QP	25	30	35	40	25	30	35	40	25	30	35	40
balloons	33.52	33.23	32.60	31.54	−0.63	−0.59	−0.51	−0.35	0.46	0.43	0.28	0.13
kendo	35.90	35.45	34.57	33.22	−1.30	−1.21	−1.03	−0.85	0.41	0.30	0.17	0.11
newspaper	26.10	26.06	25.95	25.69	−0.46	−0.47	−0.46	−0.41	0.20	0.20	0.18	0.11
poznanhall2	35.67	35.34	34.82	34.03	−1.48	−1.24	−1.01	−0.76	0.71	0.69	0.53	0.37
average of seq.	32.80	32.52	31.99	31.12	−0.97	−0.88	−0.75	−0.59	0.44	0.40	0.29	0.18
average of QPs						−0.80			0.33			

Table 3 SSIM performance over QPs

Sequence	VSRS-1D Fast	TDVS
balloons	0.9478	0.9484
kendo	0.9590	0.9593
newspaper	0.9238	0.9242
poznanhall2	0.9570	0.9582
average	0.9469	0.9475

Table 4 Percentage of a pixel using temporal-based prediction in synthesised view for different QPs

QP	25	30	35	40
balloons	8.00	6.77	3.27	3.27
kendo	11.42	10.17	6.84	3.17
newspaper	10.69	10.35	9.63	6.67
poznanhall2	5.93	6.59	4.56	3.98
average	9.01	8.47	6.08	4.27

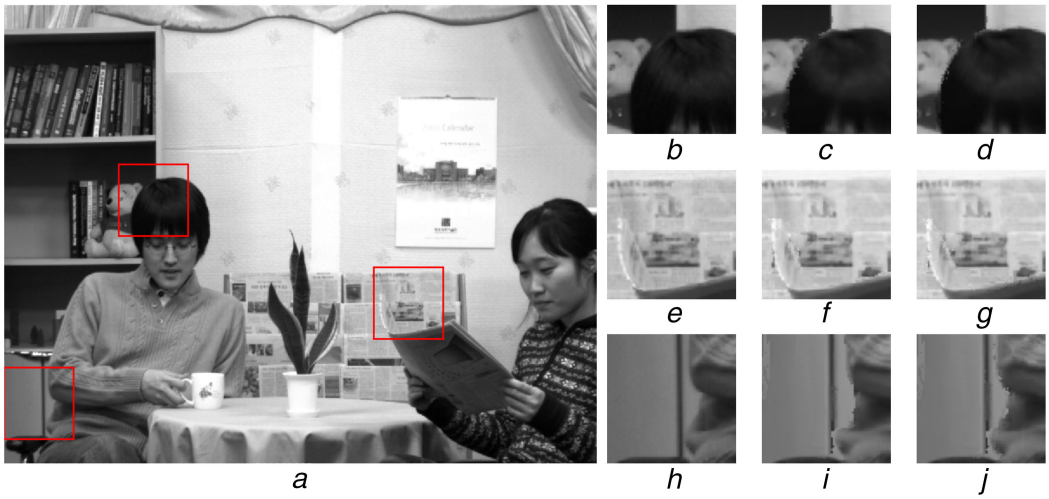

Fig. 6 *Parts of a frame synthesized with the reference VSRS-1D Fast and the proposed method.*
(a)*Rendering of a frame of sequence newspaper*
(b, e, h) QP25 with enlarged regions in the original colour
(c, f, i) Synthesised images obtained by VSRS-1D Fast
(d, g, j) Synthesised images obtained by the proposed method

Table 5 Overhead bits (%) for different QPs

SI rep.	FL				HC			
QP	25	30	35	40	25	30	35	40
balloons	9.41	15.13	8.80	4.09	4.59	7.54	5.03	2.78
kendo	9.81	8.03	4.21	2.33	4.97	4.57	2.87	1.74
newspaper	10.46	20.72	22.94	10.06	5.22	10.40	12.35	6.54
poznanhall2	14.00	31.13	26.96	16.09	8.35	18.61	15.43	10.28
average	10.92	18.75	15.73	8.14	5.78	10.28	8.92	5.33

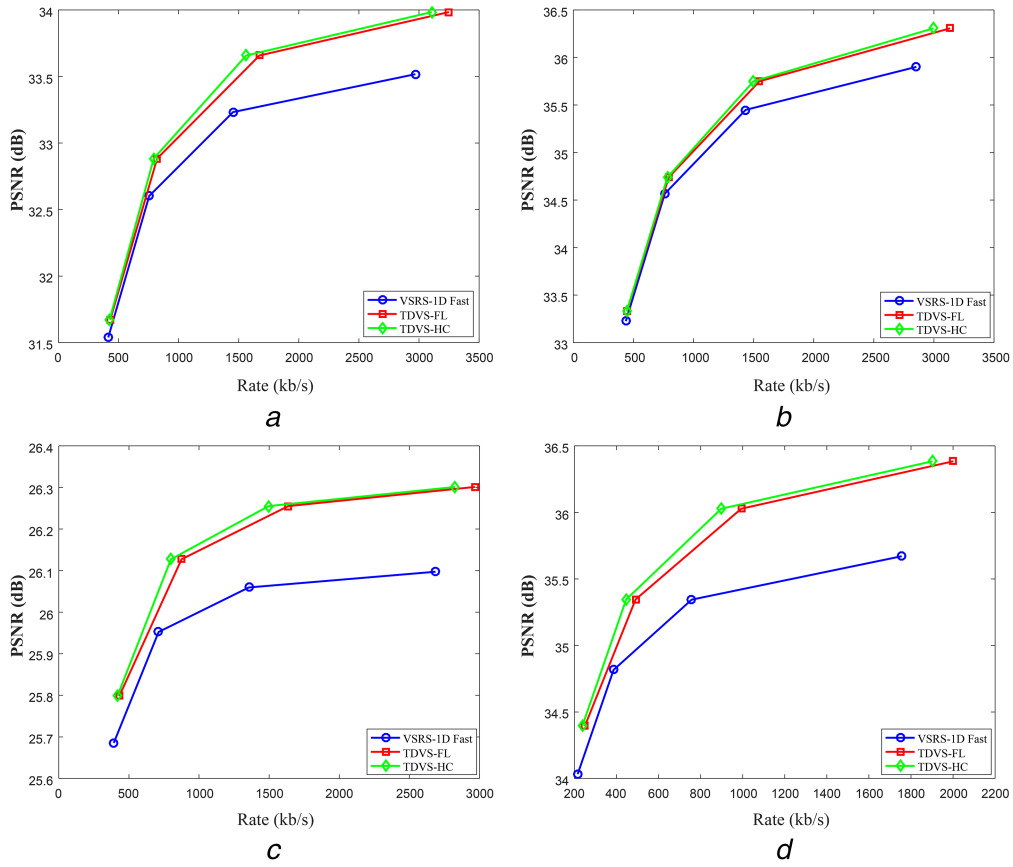

Fig. 7 *RD performance*

Table 6 Δ PSNR (dB) improvement of TDVS

SI representation	FL	HC
balloons	0.25	0.29
kendo	0.17	0.20
newspaper	0.14	0.16
poznanhall2	0.41	0.49
average	0.24	0.28

Table 7 Average lengths of sub-arrays for different QPs

QP	25	30	35	40
balloons	132.11	142.14	313.22	905.43
kendo	188.21	407.90	991.66	1479.08
newspaper	175.49	170.02	272.28	784.79
poznanhall2	258.46	299.49	469.97	1225.64
average	188.57	254.89	511.78	1098.74

6 References

[1] Smolic, A., Mueller, K., Merkle, P., *et al.*: '3D video and free viewpoint video- technologies, applications and MPEG standards'. 2006 IEEE Int. Conf. on Multimedia and Expo, Toronto, Canada, 2006, pp. 2161–2164

[2] Smolic, A., Mueller, K., Merkle, P., *et al.*: 'Multi-view video plus depth (MVD) format for advanced 3D video systems', ISO/IEC JTC1/SC29/WG11 and ITU-T SG16 Q, 2007, 6, p. 2127

[3] Fehn, C.: 'Depth-image-based rendering (DIBR), compression, and transmission for a new approach on 3D-TV', Proc. SPIE Electronic Engineering, San Jose, USA, 2004, pp. 93–105

[4] Yang, L., Yendo, T., Tehrani, M.P., *et al.*: 'Artifact reduction using reliability reasoning for image generation of FTV', *J. Vis. Commun. Image Represent.*, 2010, **21**, (5–6), pp. 542–560

[5] Lee, C., Ho, Y.S.: 'Boundary filtering on synthesized views of 3D video'. Second Int. Conf. on Future Generation Communication and Networking Symposia, Sanya, China, 2008, pp. 15–18

[6] Mueller, K., Smolic, A., Dix, K., *et al.*: 'View synthesis for advanced 3D video systems', *EURASIP J. Image Video Process.*, 2009, **2008**, (1), p. 438148

[7] Criminisi, A., Pérez, P., Toyama, K.: 'Region filling and object removal by exemplar-based image inpainting', *IEEE Trans. Image Process.*, 2004, **13**, (9), pp. 1200–1212

[8] Guillemot, C., Le Meur, O.: 'Image inpainting: overview and recent advances', *IEEE Signal Process. Mag.*, 2014, **31**, (1), pp. 127–144

[9] Bertalmio, M., Sapiro, G., Caselles, V., *et al.*: 'Image inpainting'. Proc. 27th Annual Conf. on Computer Graphics and Interactive Techniques, New Orleans, USA, 2000, pp. 417–424

[10] Daribo, I., Pesquet-Popescu, B.: 'Depth-aided image inpainting for novel view synthesis'. 2010 IEEE Int. Workshop on Multimedia Signal Processing (MMSP), Saint Malo, France, 2010, pp. 167–170

[11] Purica, A.I., Mora, E.G., Pesquet-Popescu, B., *et al.*: 'Improved view synthesis by motion warping and temporal hole filling'. 2015 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), Brisbane, Australia, 2015, pp. 1191–1195

[12] Chen, K.Y., Tsung, P.K., Lin, P.C., *et al.*: 'Hybrid motion/depth-oriented inpainting for virtual view synthesis in multiview applications'. 2010 3DTV Conf.: The True Vision –Capture, Transmission and Display of 3D Video, Tampere, Finland, 2010, pp. 1–4

[13] Purica, A.I., Cagnazzo, M., Pesquet-Popescu, B., *et al.*: 'View synthesis based on temporal prediction via warped motion vector fields'. 2016 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, 2016, pp. 1150–1154

[14] Purica, A.I., Mora, E.G., Pesquet-Popescu, B., *et al.*: 'Multiview plus depth video coding with temporal prediction view synthesis', *IEEE Trans. Circuits Syst. Video Technol.*, 2016, **26**, (2), pp. 360–374

[15] Le Dinh, M., Tung, L.V., Van, X.H., *et al.*: 'Improving 3D-TV view synthesis using motion compensated temporal interpolation'. 2016 Int. Conf. on Advanced Technologies for Communications (ATC), Hanoi, Vietnam, 2016, pp. 312–317

[16] Liu, C., Freeman, W.T., Adelson, E.H., *et al.*: 'Beyond pixels: exploring new representations and applications for motion analysis'. PhD thesis, Massachusetts Institute of Technology, 2009

[17] Sze, V., Budagavi, M., Sullivan, G.J.: 'High efficiency video coding (HEVC)', in '*Integrated circuit and systems, algorithms and architectures*' (Springer, Berlin Heidelberg, 2014), **39**, p. 40

[18] Kim, I.K., Min, J., Lee, T., *et al.*: 'Block partitioning structure in the HEVC standard', *IEEE Trans. Circuits Syst. Video Technol.*, 2012, **22**, (12), pp. 1697–1706

[19] Zhu, C., Zhao, Y., Yu, L., *et al.*: '*3D-TV system with depth-image-based rendering*' (Springer-Verlag, New York, 2014)

[20] Di Stefano, L., Bulgarelli, A.: 'A simple and efficient connected components labeling algorithm'. Proc. Int. Conf. on Image Analysis and Processing, 1999, Venice, Italy, 1999, pp. 322–327

[21] Sullivan, G.J., Wiegand, T.: 'Rate-distortion optimization for video compression', *IEEE Signal Process. Mag.*, 1998, **15**, (6), pp. 74–90

[22] Stockhammer, T., Kontopodis, D., Wiegand, T.: 'Rate-distortion optimization for JVT/H. 26L video coding in packet loss environment', 2002

[23] Wiegand, T., Girod, B.: 'Lagrange multiplier selection in hybrid video coder control'. Proc. 2001 Int. Conf. on Image Processing, Thessaloniki, Greece, 2001, pp. 542–545

[24] Chen, Y., Tech, G., Wegner, K., *et al.*: 'Test model 11 of 3D-HEVC and MV-HEVC', JCT-3V Doc., JCT3V-J1003, Geneva, CH, 2015

[25] Rusanovskyy, D., Müller, K., Vetro, A.: 'Common test conditions of 3DV core experiments', ITU-T SG, 2013, p. 16

[26] Sullivan, G.J., Ohm, J., Han, W.J., *et al.*: 'Overview of the high efficiency video coding (HEVC) standard', *IEEE Trans. Circuits Syst. Video Technol.*, 2012, **22**, (12), pp. 1649–1668

[27] Hore, A., Ziou, D.: 'Image quality metrics: PSNR vs. SSIM'. 2010 20th Int. Conf. on Pattern Recognition (ICPR), Istanbul, Turkey, 2010, pp. 2366–2369

[28] Bjontegarrd, G.: 'Calculation of average PSNR differences between RD-curves', VCEG-M33, 2001