Virtual-View-Assisted Video Super-Resolution and Enhancement

Zhi Jin, *Student Member, IEEE*, Tammam Tillo, *Senior Member, IEEE*, Chao Yao, Jimin Xiao, *Member, IEEE*, and Yao Zhao, *Senior Member, IEEE*

***Abstract*— A 3-D multiview video gives users an experience that is different from that provided by a traditional video; however, it puts a huge burden on limited bandwidth resources. Mixed-resolution video in a multiview system can alleviate this problem by using different video resolutions for different views. However, to reduce visual uncomfortableness and to make this video format more suitable for free-viewpoint television, the low-resolution (LR) views need to be super-resolved to the target full resolution. In this paper, we propose a virtual-view-assisted super-resolution algorithm, where the inter-view similarity is used to determine whether the missing pixels in the super-resolved frame need to be filled by virtual-view pixels or by spatial interpolated pixels. The decision mechanism is steered by the texture characteristics of the neighbors of each missing pixel. Furthermore, the inter-view similarity is used, on the one hand, to enhance the quality of the virtual-view-copied pixels by compensating the luminance difference between different views and, on the other hand, to enhance the original LR pixels in the super-resolved frame by reducing their compression distortion. Thus, the proposed method can recover the details in regions with edges while maintaining good quality at smooth areas by properly exploiting the high-quality virtual-view pixels and the directional correlation of pixels. The experimental results demonstrate the effectiveness of the proposed approach with a peak signal-to-noise ratio gain of up to 3.85 dB.**

***Index Terms*— Interpolation-based super-resolution (SR), mixed resolution (MR) system, multiview video, virtual view.**

I. Introduction

WITH the development of video technology, 3-D video applications are increasingly accessible to customers. Among these applications, 3-D television (3-DTV) [1] and free-viewpoint television (FTV) [2] have attracted a lot of attention. 3-DTV provides viewers the perception of real-world scenes using multiple views at different viewpoints. FTV allows viewers to freely choose any viewpoint in the scene within a certain range. The interactive and vivid experience of realistic scenes via 3-D video relies on a huge amount of texture and depth map data. Hence, it puts pressure on the acquisition, storage, and transmission processes, especially for limited bandwidth applications [3]. One effective solution, for such kind of problem, has been proposed in [4] and [5] that uses mixed-resolution (MR) video, in which at least one of the views is captured at a low resolution (LR), while the others are captured at a full resolution (FR). The MR video in comparison with FR video significantly reduces the amount of captured, transmitted, stored data, and processing time which is the bottleneck of real-time applications. Nevertheless, to meet the requirements of high definition, to reduce visual uncomfortableness, and to make the video format more suitable for FTV, the LR video needs to be super-resolved to FR size using the super-resolution (SR) techniques at the decoder side [4]. Therefore, in an MR video system, the final quality will depend on the performance of the SR algorithm.

In general, the image SR algorithms can be classified into three categories: 1) multi-image-based SR algorithms [6], [7]; 2) example-learning-based SR algorithms [8], [9]; and 3) interpolation-based SR algorithms [10], [11]. The multi-image-based SR approaches are based on the assumption that the lost high frequency in an LR image can be recovered through multiple LR images of the same scene with subpixel misalignments. These approaches can be realized on both frequency and spatial domains [6], [7]. However, this kind of method highly relies on the choice of the regularization parameters and the number of LR images, which are not easy to be obtained in reality [12]. In contrast, the example-learning-based SR approaches assume that it is possible to predict the missing high-frequency details in a single LR image by a group of LR and FR image pairs [13]. Unfortunately, their performance largely depends on the choice of training samples, so unsuitable training samples produce some obvious artifacts in the recovered high-resolution (HR) image [12]. To tackle this dependency problem on learning or training data and to improve the practicality of SR algorithm, especially for real-time systems, interpolation-based SR algorithms have been proposed. The spirit of the interpolation-based SR algorithms is that the missing HR pixels can be estimated using the information from neighboring LR pixels. Compared with the previous two kinds of SR methods, interpolation-based SR methods gain their popularity in real-time application mainly due to their computational simplicity. However, the main drawback

Manuscript received May 30, 2014; revised December 31, 2014; accepted March 3, 2015. Date of publication March 13, 2015; date of current version March 3, 2016. This work was supported in part by the National Natural Science Foundation of China under Grant 61210006 and Grant 60972085, in part by the 973 Program under Grant 2012CB316400, and in part by the Program for Changjiang Scholars and Innovative Research Team in University under Grant IRT201206. This paper was recommended by Associate Editor E. Cetin.

Z. Jin, T. Tillo, and J. Xiao are with the Department of Electrical and Electronic Engineering, Xi'an Jiaotong-Liverpool University, Suzhou 215123, China (e-mail: tammam.tillo@xjtlu.edu.cn).

C. Yao and Y. Zhao are with the Institute of Information Science, Beijing Jiaotong University, Beijing 100044, China, and also with the Beijing Key Laboratory of Advanced Information Science and Network Technology, Beijing 100044, China.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCSVT.2015.2412791

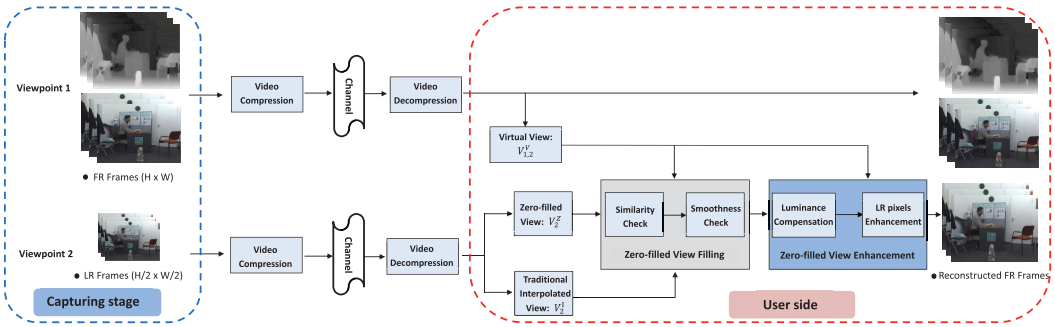


Fig. 1. Framework of the proposed SR method.

of these methods is their inability to fully exploit the scene content during the interpolation process and, consequently, they are prone to blur high-frequency details (edges). To overcome some of these weaknesses, Zhang and Wu [10] proposed to adaptively fuse the LR pixels on the two diagonal directions according to the linear minimum mean-square error estimation technique. Li and Nguyen [11] proposed using the edge-direction information implicitly in the interpolation process, with the aid of Markov random field (MRF) model. Although these edge-guided algorithms can preserve the edge structure, the recovered high-frequency details are still limited. Therefore, Garcia *et al.* [14] proposed to use the high-frequency content from neighboring FR views and the corresponding depth information to recover the high-frequency content in the LR view. In [15], the 3-D video SR method relays on the fusion of the weighted sum of nonlocal patches, and FR view to LR view mapping. Different from previous approaches, which focus solely on using spatial information, in [16], a 3-D MRF model was used to find the optimized patches from a database of HR images both in spatial and temporal domains to super-resolved the LR frames.

In this paper, we propose a new virtual-view-assisted SR and enhancement algorithm, where the exploitation of the virtual-view information and the interpolated frames has two benefits, first, the high-frequency information contained in FR views can be properly utilized to super-resolve LR views; second, the inter-view redundancy will be used to enhance the original LR pixels in the super-resolved views and to compensate the luminance difference between views. The experimental results have shown that the proposed algorithm achieves superior performance with respect to the interpolation-based algorithms.

The rest of this paper is arranged as follows. Details of the proposed SR method will be introduced in Section II. Several algorithms for choosing the thresholds are given in Section III. The generalization of proposed method is presented in Section IV, and the experimental results are presented in Section V. Section VI concludes this paper and also discusses our future work.

II. Proposed Super-Resolution Method

In [17] and [18], it has been shown that a comfortable viewing of MR format could be achieved when the resolution of FR view is twice as much the resolution of the LR view in both horizontal and vertical directions, whereas, higher ratios of the FR to LR resolutions will result in unacceptable subjective quality. In the following, this ratio will be dubbed *resolution factor* for brevity. This paper will only address a resolution factor of 2 by using the findings in [17] and [18], and the framework of the proposed virtual-view-assisted interpolation-based SR algorithm is shown in Fig. 1. At viewpoint 1, the FR textures and associated depth maps with frame size $H \times W$ are compressed and transmitted to the receiver side. Meanwhile, the texture at viewpoint 2 has half resolution of the FR view in both horizontal and vertical directions. The FR decoded textures and depth maps will be denoted by $\boldsymbol{V}_1^F$ and $\mathbf{D}_1^F$, respectively, while the decoded LR texture sequence will be denoted by $\boldsymbol{V}_2^L$. At the decoder side, the decompressed LR view is used to generate two intermediate FR versions at viewpoint 2. The first version ($\boldsymbol{V}_2^I$) is obtained using an interpolation method, such as bilinear or bicubic. The second version ($\boldsymbol{V}_2^Z$) is the zero-fill version of the LR view, where the original LR samples, placed at positions with indices $(2i-1,\ 2j-1)$, are separated by inserted zeros. This version will be used as basis to generate the final super-resolved FR version at viewpoint 2. The zero-inserted positions in this frame will be replaced by pixels from either the interpolated view, $\boldsymbol{V}_2^I$, or from the FR-view-generated virtual view at viewpoint 2 using the 1-D depth image-based rendering (DIBR) process [19] from one reference view to another view without any postprocessing (i.e., no hole filling). This virtual view will be referred to $\boldsymbol{V}_{1,2}^V$ in the following sections. Deciding which pixels to use to replace the inserted zeros will be driven by a similarity and smoothness check mechanism which will be explained in Section II-A. To further improve the quality of the super-resolved FR frames at viewpoint 2, some enhancement methods will be proposed in Section II-B.

A. Zero-Filled View Filling

To generate an FR frame from the corresponding LR version and recover most of the lost high-frequency information in the capturing stage, both the virtual view and the interpolated frame are used as candidates in this paper. Since the virtual view is only synthesized from the neighboring FR view, the inter-view redundancy and the high-frequency component of the FR frame can be exploited in the proposed SR approach. However, the virtual view might be affected by some holes and cracks due to the wrapping process and inaccurate depth map. Therefore, the similarity between the original LR pixels

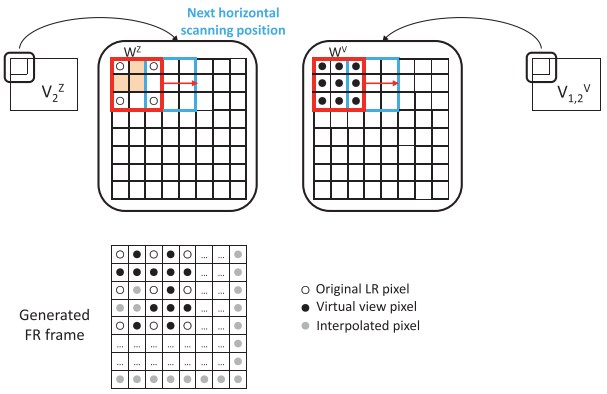


Fig. 2. Pictorial representation of the similarity check process and the generation of FR frame.

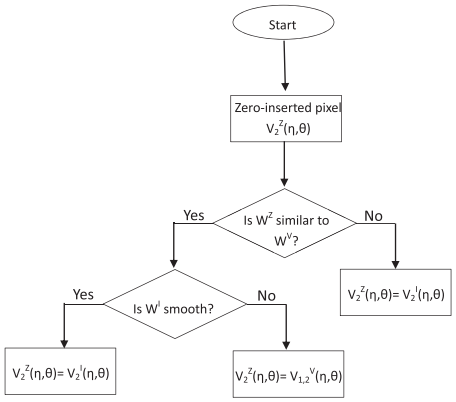


Fig. 3. Flowchart of the zero-filled view filling stage.

in $\boldsymbol{V}_2^Z$ and the corresponding pixels in the virtual view is measured to ensure that only proper virtual-view pixels are selected to replace the zero-filled pixels in $\boldsymbol{V}_2^Z$. This process will minimize the probability of copying some holes from the virtual view into $\boldsymbol{V}_2^Z$. The similarity check mechanism consists of two 3×3 scanning windows W^Z and W^V which synchronously scan the zero-filled view and the virtual view, respectively. A pictorial representation of this process is shown in Fig. 2. The centers of these windows are used as the origin of their coordinate systems, thus, for example, $W^Z(-1,-1)$ stands for the upper left corner pixel in the window W^Z. The two windows move in a raster-scan mode by sliding two pixels at a time, so as to be always centered at the zero-filled pixels, i.e., $(2i, 2j)$ with $1 \le i \le H/2$ and $1 \le j \le W/2$. This ensures that for the zero-filled view, there are four LR pixels at the corners of the window W^Z to measure the local texture similarity between the zero-filled view and the virtual view. In this paper, the sum of absolute difference (SAD) is used for this purpose as[1]

$$D_{\mathrm{SC}} = \sum_{\eta\in\{-1,1\};\theta\in\{-1,1\}} |W^Z(\eta,\theta) - W^V(\eta,\theta)|. \quad (1)$$

In this case, a hole due to DIBR process in any corner of W^V will lead, in general, to a large SAD value. Therefore, this will be used as an indication that the local virtual-view pixels in the current window, W^V, are not appropriate for filling the corresponding zero positions in the zero-filled view. Consequently, the zero-filled pixel $(2i, 2j)$ and its two causal neighbors, i.e., $(2i-1, 2j)$ and $(2i, 2j-1)$, will be filled by the corresponding interpolated pixels from $\boldsymbol{V}_2^I$ if the SAD value is larger than a threshold T_{si}, as shown in

$$V_2^Z(\eta,\theta) = V_2^I(\eta,\theta); \quad D_{\mathrm{SC}} \ge T_{\mathrm{si}} \quad (2)$$

where $(\eta,\theta) \in \mathcal{C}$ and $\mathcal{C} = \{(2i,2j),(2i-1,2j),(2i,2j-1)\}$. Hence, except for pixel-size holes located at the zero-filled positions, this mechanism will minimize the possibility of mistakenly copying some hole pixels from W^V into W^Z.

For the case when the SAD measure is smaller than T_{si}, which indicates that the diagonal pixels in the two windows are relatively similar, a further check is carried out to determine the proper approach to fill the zero-filled positions in V_2^Z. If the area encompassing W^Z is smooth then interpolation algorithms could be better than the virtual view to estimate the zero-filled pixels. This is because chromatic discrepancies among different viewpoints make the obtained virtual-view pixels less accurate than the interpolated pixels to represent the missed information for the smooth areas. The chromatic discrepancies phenomenon happens due to the scene illumination difference, camera calibration, and jitter speed, even if the capturing cameras have been adjusted to the same configuration [20]. Hence, based on this fact, zero-filled pixels in smooth areas will be replaced by their counterparts from $\boldsymbol{V}_2^I$. On the other hand, for nonsmooth areas, such as edges, interpolation algorithms intrinsically fail to estimate proper values for the zero-filled pixels, whereas, the virtual view generated from the FR view carries significant amount of information related to those nonsmooth areas. Thus, for this kind of areas, the zero-filled pixels will be replaced by their counterparts in the virtual view $\boldsymbol{V}_{1,2}^V$.

The previous paradigm is implemented in the second step, where the smoothness of a 3×3 area, W^I, centered at the pixel $(2i, 2j)$ in $\boldsymbol{V}_2^I$ is checked; in this paper, this has been done by measuring the standard deviation, σ^s. The motivation behind using the window W^I to measure the local smoothness is that a nontrivial interpolator uses more than eight-connected neighbors in the estimation process to preserve the local regularity [21]. Consequently, the five estimated pixels along with the four corners of W^I carry more information about the local smoothness of the area, than the four LR pixels at the corners of the W^Z window. The outcomes of the smoothness check stage could be summarized in

$$V_2^Z(\eta,\theta) = \begin{cases} V_2^I(\eta,\theta); & \sigma^s < T_{\mathrm{sm}} \\ V_{1,2}^V(\eta,\theta); & \sigma^s \ge T_{\mathrm{sm}} \end{cases} \quad (3)$$

where $(\eta,\theta) \in \mathcal{C}$. In (3), T_{sm} is a threshold to determine whether an area surrounding the pixel $(2i, 2j)$ has smooth or nonsmooth texture. A flowchart of similarity check and smoothness check stages is shown in Fig. 3. As for the boundary pixels, they are copied from the interpolated view directly.

[1]The SAD and Euclidean distance in this case will almost lead to the same results.

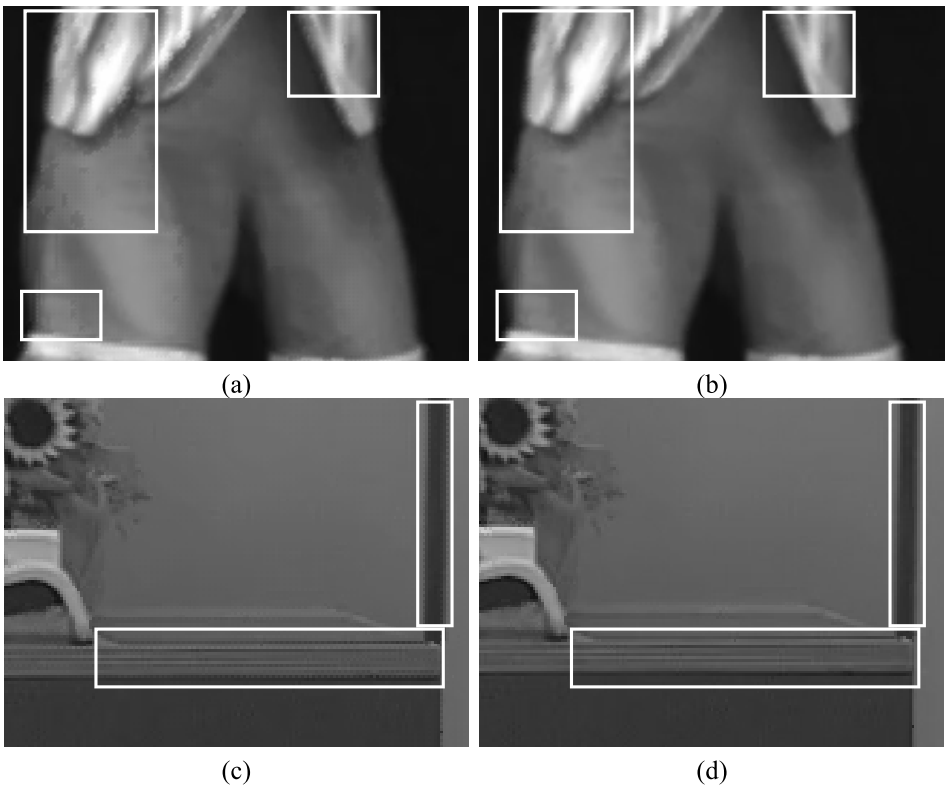


Fig. 4. Comparison of the effect of luminance compensation on the first frame of *Pantomime* and *Bookarrival* sequences. The two images on the left show the artifacts in the super-resolved frames without luminance compensation and the two images on the right show the visual effects of same frame but after luminance compensation [better perception could be achieved by viewing the images at their full resolution, which are (a) and (b) 620×775 and (c) and (d) 620×884].

B. Zero-Filled View Enhancement

In the previous stage, all the zero-filled positions will be filled by either virtual-view pixels or interpolated pixels. However, the recovered FR frame will be affected by compression distortion, virtual-view-introduced distortion, and interpolation-induced distortion, therefore, in this paper, two methods are proposed to reduce the overall distortion, and enhance the final quality of the generated FR view.

1) Luminance Compensation: In real video capturing scenarios, different views will have slightly different luminance. This is due to some inevitable factors, such as imbalanced light condition of the scene, inaccurate synchronization of cameras, which cause some projection differences among different viewpoints. In addition, the LR and FR views will have different quality after compression, especially at large quantization parameters (QP), this is demonstrated in Table I for four sequences. This is because the LR view has more details in each macroblock than the FR view, thus even with the same QP, the quality of its compressed version will be lower than its counterpart in the FR view. All of these factors cause some jagged edges of the reconstructed FR frames when using the virtual view to recover the zero-filled positions in $\boldsymbol{V}_2^Z$, and an example of this artifact is shown in the highlighted areas in Fig. 4(a) and (c). Hence, a luminance compensation mechanism is proposed, which adjusts the brightness of the copied pixels from the virtual view into $\boldsymbol{V}_2^Z$ (i.e., the virtual-view-based recovered pixels) to have harmonious brightness with the surrounding LR pixels. If the set of pixels $\mathcal{C}$ of current pixel $(2i, 2j)$ are recovered from the virtual view, the average luminance difference between the two sliding windows W^Z and W^V centered at $(2i, 2j)$ will be evaluated as

$$D_{\mathrm{LC}} = \frac{1}{4} \sum_{\eta \in \{-1,1\}; \theta \in \{-1,1\}} [W^Z(\eta, \theta) - W^V(\eta, \theta)]. \quad (4)$$

If the absolute value of D_{LC} is larger than a threshold T_l, the compensation process will be used to update the intensity of the pixels $\mathcal{C}$. The reason behind using the threshold T_l is to eliminate the effect of compression distortion on the luminance compensation process. In fact, given that a

TABLE I
PSNR DIFFERENCES BETWEEN FR AND LR VIEWS USING H.264 FOR (a) *Bookarrival*, (b) *Doorflower*, (c) *Laptop*, AND (d) *Champagne* WITH QP = 22, 27, 32, 37, 42, 47

Seq.	Bookarrival	Doorflower	Laptop	Champagne
QP	ΔPSNR(dB)	ΔPSNR(dB)	ΔPSNR(dB)	ΔPSNR(dB)
22	1.00	0.78	0.83	0.65
27	1.48	1.15	1.22	1.28
32	2.07	1.88	1.82	2.23
37	2.55	2.42	2.31	2.75
42	2.69	2.68	2.56	2.80
47	2.22	2.20	2.19	3.05

small number of pixels are used to estimate D_{LC}, it will be highly likely that this estimated value is biased by the amount of compression distortion affecting W^Z and W^V. Nevertheless, the use of the threshold will ensure that mainly luminance differences get compensated, and the small window size will ensure that luminance compensation is performed locally. Once the luminance compensation process is invoked for a set $\mathcal{C}$, then for each of its three pixels, a proper amount of compensation will be determined using the closest available neighbors for each of the pixel in $\mathcal{C}$ to avoid blurring the edges. For example, for the pixel $(2i-1, 2j)$ its two horizontal neighbors $(2i-1, 2j-1)$ and $(2i-1, 2j+1)$ will be used to evaluate its compensation value ΔY^h

$$\Delta Y^h = \frac{1}{2}[\Delta W(-1,-1) + \Delta W(-1,+1)] \tag{5}$$

where $\Delta W(\eta,\theta) = W^Z(\eta,\theta) - W^V(\eta,\theta)$. Now, the position $(2i-1, 2j)$ in the zero-filled view will be filled by $V^V_{1,2}(2i-1, 2j) + \Delta Y^h$ instead of $V^V_{1,2}(2i-1, 2j)$. Similarly, the compensation value for the pixel $(2i, 2j-1)$ will be computed starting from its two vertical neighbors as $\Delta Y^v = 1/2[\Delta W(-1,-1) + \Delta W(+1,-1)]$ and consequently $V^Z_2(2i, 2j-1) = V^V_{1,2}(2i, 2j-1) + \Delta Y^v$. As for the center pixel, $(2i, 2j)$, it will be updated as $V^Z_2(2i, 2j) = V^V_{1,2}(2i, 2j) + \Delta Y^c$. However, given that the pixel $(2i, 2j)$ is at equal distance from the four corners, its compensation value will be evaluated as

$$\Delta Y^c = \frac{1}{4}[\Delta W(-1,-1) + \Delta W(-1,+1) + \Delta W(+1,-1) + \Delta W(+1,+1)]. \tag{6}$$

Some luminance compensation results are shown in Fig. 4(b) and (d).

2) LR Pixels Enhancement in the Super-Resolved View: The previous section proposes an enhancement mechanism for the virtual-view-based recovered pixels, whereas, in this section, a mechanism to enhance the quality of the other pixels in $\boldsymbol{V}^Z_2$, i.e., the original LR pixels, is proposed. This is particularly important given that these pixels suffer from more compression distortion than their counterparts in the FR view, as shown in Table I. The proposed enhancement method in this section exploits the inter-view correlation to achieve its objective. In fact, in multiview system, adjacent views have large similarity, so the same content may appear in two different positions in the two adjacent views. Hence, if the same content is separately encoded in the two views, then their compression distortions could be partially canceled out. To show why the proposed mechanism improves the performance, and to show its principle, let us consider a point in the scene $\hat{v}$ which is viewed from two viewing points, which means it is not occluded in any of these two views. Let us denote the projection of this point into viewpoints 1 and 2 by $\hat{V}^F_1(\mu,\nu)$ and $\hat{V}^L_2(i,j)$, respectively. Apart from small differences, due to the nature and relative position of the lighting source in the scene, the previous two values could be regarded as similar, i.e., $\hat{V}^F_1(\mu,\nu) \approx \hat{V}^L_2(i,j) = \hat{v}$, the smaller the baseline is the more correct this assumption is.[2] Hence, in the following sections, we will assume that $\hat{V}^F_1(\mu,\nu) = \hat{V}^L_2(i,j) = \hat{v}$. These two projections will be compressed separately in the two viewpoints, so they become: 1) $V^F_1(\mu,\nu) = \hat{v} + d_1$ and 2) $V^L_2(i,j) = \hat{v} + d_2$, where d_1 and d_2 are the distortion caused by video compression on views 1 and 2, respectively. Since the compression can be treated as a random process with mean value being $E\{d_1\} = E\{d_2\} = 0$, the variance of the distortion affecting views 1 and 2 will be $\sigma_1^2 = E\{d_1^2\}$ and $\sigma_2^2 = E\{d_2^2\}$, respectively. Then at the decoder side, and as explained previously, the zero-filled view is obtained from the LR view by inserting zeros in between its pixels, thus

$$V^Z_2(2i-1, 2j-1) = V^L_2(i,j) = \hat{v} + d_2. \tag{7}$$

At this point, let us assume that the wrapping process works accurately and maps $V^F_1(\mu,\nu)$ into $V^V_{1,2}(2i-1, 2j-1)$ without introducing tangible wrapping distortion. This assumption implies that the depth information is accurate, in this case

$$V^V_{1,2}(2i-1, 2j-1) = \hat{v} + d_1. \tag{8}$$

If at the decoder side, the pixel at position $(2i-1, 2j-1)$ in the zero-filled view is replaced by the average of $V^Z_2(2i-1, 2j-1)$ and $V^V_{1,2}(2i-1, 2j-1)$ then the expected compression distortion could be evaluated using (7) and (8) as

$$\sigma^2 = E\left\{\left(\frac{\hat{v} + d_2 + \hat{v} + d_1}{2} - \hat{v}\right)^2\right\} = E\left\{\left(\frac{d_2 + d_1}{2}\right)^2\right\}. \tag{9}$$

Since views 1 and 2 are separately compressed, $E\{d_1 d_2\} = 0$. In the general case, $d_1 \le d_2$ even when using the same QP for the LR and FR views, consequently, $(\sigma_2^2/4) \le \sigma^2 \le (\sigma_2^2/2)$. This means that the equivalent distortion of the pixels at $(2i-1, 2j-1)$, where $1 \le i \le H/2$ and $1 \le j \le W/2$ in the zero-filled view will be reduced.

It is worth noting that the averaging process can only be applied to those pixels in $\boldsymbol{V}^Z_2$ which have equivalent pixels in $\boldsymbol{V}^V_{1,2}$, thus holes and occluded areas need to be excluded from this process, so to ensure this, the similarity and smoothness check mechanism proposed in Section II-A will be used as well here. Since the sliding window used in this process moves in a raster scan fashion, then except for some border pixels, each LR pixel will appear in four different windows. Therefore, only when the LR pixel is regarded similar to its counterpart virtual-view pixel in four measurements, then it will be replaced by $[V^Z_2(2i-1, 2j-1) + V^V_{1,2}(2i-1, 2j-1)]/2$.

III. Thresholds Evaluation

In the proposed SR approach, both the virtual and interpolated views are utilized to generate the FR frames, and two postprocessing enhancement operations are exploited to further improve the quality of the generated FR view. In this whole

[2] Although the coordinate system in the two views are related, however, they are different due to the fact that the two views have different resolutions.

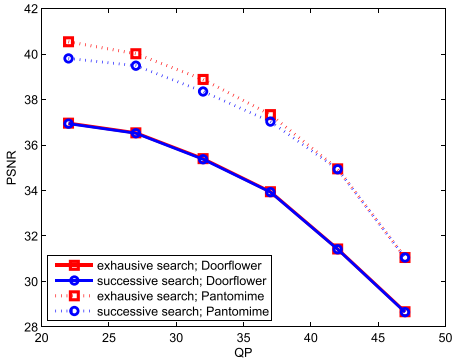


Fig. 5. Comparison of exhaustive and successive approaches for threshold determination on *Doorflower* and *Pantomime* sequences.

process, three thresholds are required. An experimental-based approach to determine the values of these thresholds could be used at the encoder side using an analysis-by-synthesis approach. Since these three thresholds are intertwined, the choice of one will have some impact on the others. Therefore, to obtain the best thresholds, the encoder needs to test different combinations of them using three nest-like loops, and then send values to the decoder. If the complexity of estimating one threshold is $\mathcal{O}(n)$, then the complexity of this exhaustive approach is $\mathcal{O}(n^3)$. A simplified approach is proposed, where the value for each threshold is obtained in a successive approach and the complexity can be consequently reduced from $\mathcal{O}(n^3)$ to $\mathcal{O}(3n)$. Some experiments were conducted on *Doorflower* and *Pantomime* sequences at QP $= \{22, 27, 32, 37, 42, 47\}$ to compare the performance of exhaustive and successive approaches and the corresponding results are shown in Fig. 5.[3] The results indicate that the proposed simplified approach can significantly reduce the complexity without large quality degradation and also indicate that although these three thresholds are intertwined, there are some other factors that have more influence on their values. It is shown in the Appendix that T_{si} and T_l could be represented by

$$T_{\text{si}} = \alpha\sqrt{\sigma_1^2 + \sigma_2^2} \tag{10}$$

$$T_l = \beta\sqrt{\sigma_1^2 + \sigma_2^2} \tag{11}$$

where σ_1 and σ_2 are the standard deviations of the compression distortion affecting view 1 (FR view) and view 2 (LR view), respectively. α and β are the two parameters, which depend on the sequence content.

In the successive approach, the three thresholds (or equivalently, α, β, and T_{sm}) could be either determined at the encoder frame by frame and sent to the decoder side, or just determined for the first frame and then applied on the following frames. These two approaches have been tested and the results are shown in Fig. 6. In addition, a third approach that uses a user-defined value for both α and β and the corresponding results are also shown in Fig. 6. In this approach, it is reasonable to assume α and β are larger than three based on Chebyshev's inequality. From Fig. 6, it is obvious that the successive-search approach which estimates α, β, and T_{sm} based on the first frame, and then use these values for the following frames is almost as good as the frame by frame approach, consequently, all the following experiments were conducted using this approach.

[3]Similar results, not reported here for briefness, have been obtained from other sequences.

IV. Multiview Video

The proposed virtual-view-assisted SR algorithm can also be applied to multiview multiresolution systems. Since in this kind of systems more neighboring FR views and the corresponding depth maps are available, at a given viewpoint, more virtual-view versions can be utilized. With the aid of these virtual views, the quality of the final generated FR views can be considerably improved. As shown in Fig. 7, $V_{q,k}^V$ ($q = 1, \ldots, m$ and $k = 1, \ldots, n$) is the virtual view generated from one of the adjacent FR view at viewpoint q to one of the LR view at viewpoint k. In this case, the zero-filled pixels in the zero-filled view are replaced by selecting among the available virtual views the one which better satisfies the similarity condition. Subsequently, the two enhancement methods are carried out step by step. In this way, the proposed algorithm effectively super-resolves the LR views.

V. Experimental Results

The proposed SR program will be available on http://www.mmtlab.com/download.ashx.

To objectively and subjectively evaluate the performance of the proposed algorithm, several experiments were conducted with typical 3-D video sequences, such as *Doorflower*, *Bookarrival*, *Leavelaptop*, *Pantomime*, *Champagne*, *Dog*, and *Kendo*. Since the proposed SR method is targeted for MR paradigm, and due to the lack of MR MVD sequences. We generated testing sequences by downsampled at least one of the FR views to LR for each of the tested sequences. The downsampling factor is two on both horizontal and vertical directions. The original FR views are considered as ground-truth views for the objective assessments. The DIBR technique is employed to render the virtual views, and the H.264/AVC reference software JM17.0 [22] is used to implement the coding process. The IPPP coding structure is used and one second of each sequence is tested. The QP values are $\{22, 27, 32, 37, 42, 47\}$ for both texture and depth map and for both FR and LR views. In the following experiment, the six-tap *Lanczos* interpolation filter is used as benchmark method. The other most used interpolation method *Bicubic* is also tested. Finally, peak signal-to-noise ratio (PSNR) and structural similarity index measurement (SSIM) [23] are employed to assess the objective performance.

In the experiments, we first evaluated the effectiveness of the proposed approach on stereoscopic sequences, and then compared its performance with other approaches. Second, to verify the effectiveness of each stage of the proposed method, the results of each stage are also reported. Finally, the proposed algorithm was applied on MVD sequences.

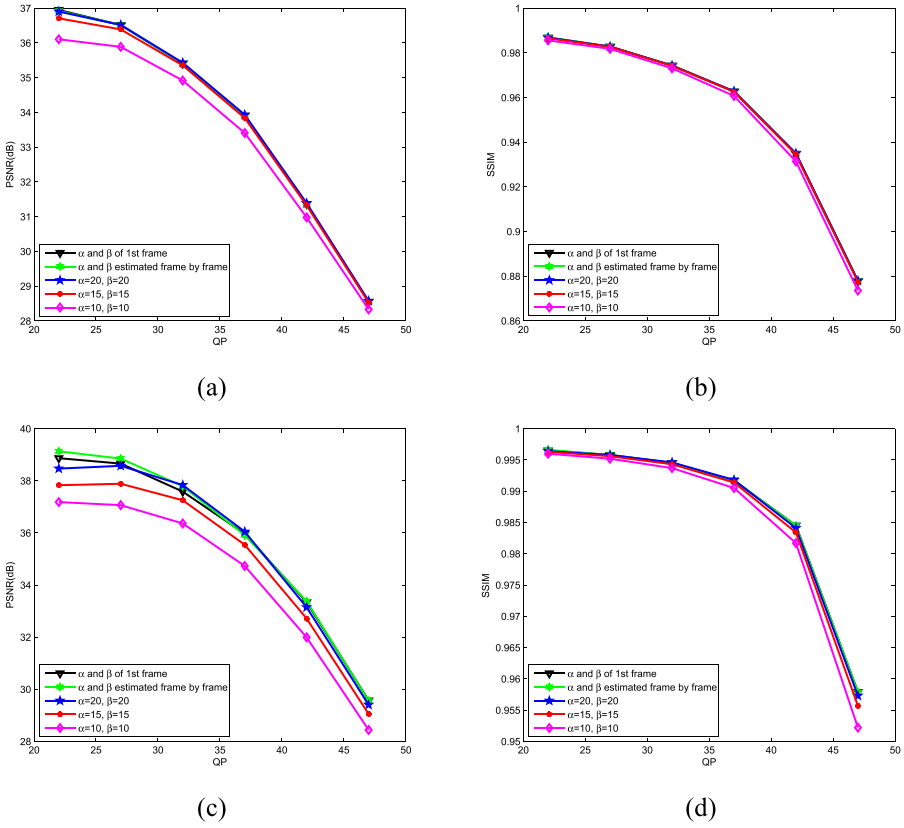


Fig. 6. PSNR and SSIM comparisons of different approaches for the evaluation of α and β. (a) and (b) Results of *Doorflower*. (c) and (d) Results of *Pantomime*.

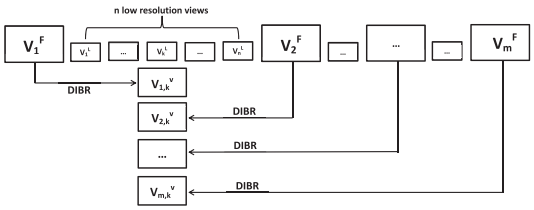


Fig. 7. Proposed algorithm for multiview multiresolution system.

TABLE II
PARAMETERS AND CHARACTERISTICS FOR EACH USED SEQUENCE

Name	Size	FR	LR	Content's Motion
Doorflower	1024 × 768	View10	View08	Moderate
Bookarrival	1024 × 768	View08	View06	Moderate
Leavelaptop	1024 × 768	View06	View07	Moderate
Pantomime	1280 × 960	View39	View40	Medium complex
Champagne	1280 × 960	View37	View38	Complex
Dog	1280 × 960	View38	View39	Medium complex
Kendo	1024 × 768	View03	View04	Complex

A. Performance Evaluation on Two-View Video

Unless otherwise noted, the characteristics and two chosen viewpoints for each testing sequence have been listed in Table II. All the PSNR and SSIM results for the luminance component of the three interpolation-based approaches are shown in Table III, where Lan = Lanczos, Bic = Bicubic, and Pro = Proposed. Results of the state-of-the-art single image SR approach via sparse coding (SC) [24] are also reported in the table, where the parameters for the publicly available code[4] are set according to [24]. It is clear that the proposed method outperforms the benchmark method and Bicubic method over all QPs both in terms of PSNR and SSIM. For most of the cases, the proposed method is also better than [24]. Table III also presents the PSNR and SSIM gains over the benchmark method which are indicated by ΔPSNR and ΔSSIM, respectively. This reveals that the PSNR gains increase with the decrease of QP values, while the SSIM gains increase with the increase of QP values. The highest PSNR gain obtained by the proposed method is 3.85 dB on the sequence *Bookarrival* when PQ = 22, while the average PSNR gain over all sequences and QPs is 2.11 dB. Although, in term of SSIM the gains are not as obvious as the PSNR ones, the SSIM gains still indicate an improvement of the objective quality compared with the benchmark method, especially when QP is very large.

To further evaluate the effectiveness of the proposed method, comparisons with the method proposed in [15] also have been carried out by adopting the same testing sequences with the same resolution and the same way to generate the testing MR sequences with a resolution factor of 2. The results of these comparisons are shown in Table IV.

In the following, we compare our proposed approach with [14]. The proposed method was tested under

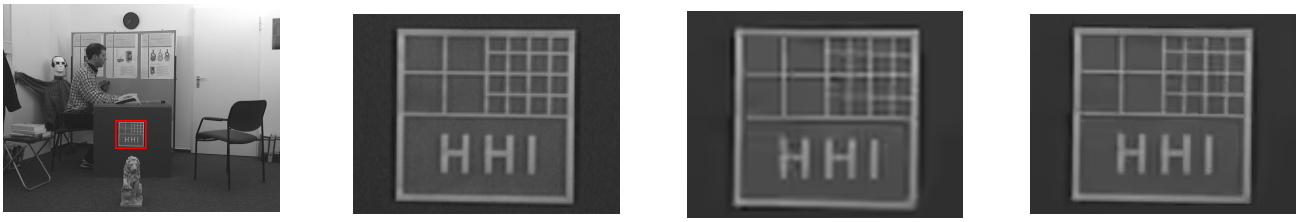
[4]http://www.ifp.illinois.edu/~jyang29/ScSR.htm

TABLE III

LUMINANCE PSNR (dB) AND SSIM RESULTS OF PROPOSED METHOD IN COMPARISON WITH OTHER METHODS AND CORRESPONDING GAINS OF THE PROPOSED METHOD OVER LANCZOS METHOD

QP	Sequence		Doorflower	Bookarrival	Leavelaptop	Pantomime	Champagne	Dog	Kendo
22	PSNR	Lan	33.12	32.98	33.31	35.50	33.95	34.96	37.56
		Cub	33.20	33.05	33.40	35.54	34.00	35.04	37.60
		SC[24]	32.83	32.57	32.94	36.31	34.79	35.43	38.05
		Pro	**36.95**	**36.83**	**36.05**	**38.87**	**36.25**	**36.05**	**39.36**
	ΔPSNR		3.83	3.85	2.74	3.37	2.30	1.09	1.80
	SSIM	Lan	0.970	0.969	0.970	0.994	0.992	0.983	0.986
		Cub	0.971	0.970	0.970	0.994	0.992	0.983	0.987
		SC[24]	0.945	0.938	0.936	0.985	0.984	0.967	0.970
		Pro	**0.987**	**0.987**	**0.984**	**0.997**	**0.995**	**0.989**	**0.990**
	ΔSSIM		0.017	0.018	0.014	0.003	0.003	0.006	0.004
27	PSNR	Lan	32.89	32.68	33.03	35.21	33.62	34.27	36.85
		Cub	32.95	32.73	33.09	35.25	33.67	34.33	36.89
		SC[24]	32.76	32.44	32.82	36.10	34.47	34.83	37.49
		Pro	**36.51**	**36.23**	**35.70**	**38.66**	**35.53**	**35.10**	**38.74**
	ΔPSNR		3.62	3.55	2.67	3.45	1.91	0.83	1.89
	SSIM	Lan	0.966	0.963	0.964	0.993	0.990	0.976	0.981
		Cub	0.967	0.963	0.964	0.993	0.990	0.976	0.981
		SC[24]	0.940	0.930	0.928	0.982	0.980	0.955	0.962
		Pro	**0.983**	**0.981**	**0.979**	**0.996**	**0.993**	**0.982**	**0.986**
	ΔSSIM		0.017	0.018	0.015	0.003	0.003	0.006	0.005
32	PSNR	Lan	32.33	32.04	32.45	34.48	32.84	32.93	35.48
		Cub	32.37	32.08	32.49	34.51	32.89	32.99	35.51
		SC[24]	32.47	32.08	32.55	35.48	33.71	33.61	36.26
		Pro	**35.41**	**35.12**	**34.81**	**37.59**	**35.00**	**33.49**	**37.45**
	ΔPSNR		3.08	3.08	2.36	3.11	2.16	0.56	1.97
	SSIM	Lan	0.957	0.951	0.953	0.991	0.985	0.961	0.972
		Cub	0.957	0.951	0.953	0.991	0.985	0.961	0.972
		SC[24]	0.931	0.916	0.917	0.977	0.972	0.937	0.950
		Pro	**0.974**	**0.972**	**0.970**	**0.995**	**0.991**	**0.967**	**0.980**
	ΔSSIM		0.017	0.021	0.017	0.004	0.006	0.006	0.008
37	PSNR	Lan	31.26	30.89	31.26	33.07	31.56	31.04	33.48
		Cub	31.30	30.92	31.30	33.12	31.62	31.10	33.51
		SC[24]	31.67	31.29	31.66	34.18	32.41	**31.89**	34.37
		Pro	**33.91**	**33.49**	**33.40**	**35.97**	**33.33**	31.36	**35.25**
	ΔPSNR		2.65	2.60	2.14	2.90	1.77	0.32	1.77
	SSIM	Lan	0.941	0.930	0.931	0.986	0.977	0.931	0.956
		Cub	0.941	0.930	0.931	0.986	0.977	0.931	0.956
		SC[24]	0.915	0.897	0.896	0.967	0.957	0.905	0.932
		Pro	**0.963**	**0.958**	**0.956**	**0.992**	**0.983**	**0.937**	**0.969**
	ΔSSIM		0.022	0.028	0.025	0.006	0.006	0.006	0.013
42	PSNR	Lan	29.46	28.99	29.49	30.74	29.50	28.42	30.87
		Cub	29.49	29.01	29.51	30.78	29.56	28.45	30.89
		SC[24]	30.35	29.92	30.41	31.95	30.55	**29.48**	31.86
		Pro	**31.36**	**31.08**	**31.31**	**33.34**	**31.05**	28.62	**32.51**
	ΔPSNR		1.90	2.09	1.82	2.60	1.55	0.20	1.64
	SSIM	Lan	0.904	0.882	0.888	0.975	0.960	0.859	0.927
		Cub	0.904	0.882	0.888	0.975	0.960	0.859	0.927
		SC[24]	0.882	0.854	0.859	0.955	0.942	0.834	0.901
		Pro	**0.935**	**0.923**	**0.924**	**0.985**	**0.971**	**0.866**	**0.949**
	ΔSSIM		0.031	0.041	0.036	0.010	0.011	0.007	0.022
47	PSNR	Lan	27.29	26.83	27.17	27.50	26.90	25.80	27.86
		Cub	27.30	26.84	27.18	27.53	26.96	25.82	27.88
		SC[24]	28.50	28.02	28.40	28.83	28.07	**27.00**	28.94
		Pro	**28.55**	**28.36**	**28.57**	**29.57**	**28.25**	25.91	**29.55**
	ΔPSNR		1.26	1.53	1.40	2.07	1.35	0.11	1.69
	SSIM	Lan	0.838	0.811	0.817	0.939	0.922	0.732	0.877
		Cub	0.838	0.811	0.817	0.939	0.922	0.732	0.878
		SC[24]	0.820	0.790	0.796	0.922	0.900	0.712	0.853
		Pro	**0.878**	**0.862**	**0.863**	**0.958**	**0.941**	**0.740**	**0.912**
	ΔSSIM		0.040	0.051	0.046	0.019	0.019	0.008	0.035

a disadvantageous condition with respect to [14], where in the latter approach the uncompressed sequence *Pantomime* and *Dog* with a resolution factor of 2 was used, and the reported gains were 2.57 and 1.06 dB over the *Lanczos* method, respectively, while, in the proposed method, even with video compression (QP $=$ 22) the gains are 3.62 and 1.22 dB, respectively. The average PSNR gain in [14] at $QP = \{22, 27, 32, 37\}$ on these two sequences



(a) (b) (c) (d)

Fig. 8. (a) Reference FR frame. (b) Cropped portion of the FR frame. The results at QP = 32 for (c) benchmark interpolation method and (d) proposed method. FR of the cropped portion is 620 × 884.

TABLE IV
PSNR (dB) AND SSIM OF SR RESULTS OBTAINED BY THE PROPOSED METHOD AND REFERENCE METHOD IN [15]

Sequence		Book	Doorflower	Laptop
PSNR	[15]	33.04	33.27	33.93
	Pro	**34.57**	**33.76**	**34.38**
ΔPSNR		1.53	0.49	0.45
SSIM	[15]	0.941	0.941	0.945
	Pro	**0.986**	**0.983**	**0.986**
ΔSSIM		0.045	0.042	0.041

TABLE V
LUMINANCE PSNR GAIN (dB) AND SSIM GAIN FOR EACH STAGE OF THE PROPOSED APPROACH; zfvf AND zfve STAND FOR ZERO-FILLED VIEW FILLING STAGE AND THE ENHANCEMENT STAGE, RESPECTIVELY

QP			22	27	32	37	42	47
Doorflower	ΔPSNR	zfvf	3.28	3.08	2.58	2.16	1.52	0.95
		zfve	0.55	0.55	0.50	0.49	0.38	0.31
	ΔSSIM	zfvf	0.016	0.017	0.018	0.023	0.034	0.044
		zfve	0.001	0.000	0.000	-0.001	-0.004	-0.005
Bookarrival	ΔPSNR	zfvf	3.04	2.81	2.44	2.06	1.68	1.27
		zfve	0.81	0.74	0.64	0.55	0.40	0.27
	ΔSSIM	zfvf	0.017	0.018	0.022	0.030	0.047	0.060
		zfve	0.001	0.000	-0.001	-0.002	-0.006	-0.009
Leavelaptop	ΔPSNR	zfvf	1.83	1.74	1.55	1.41	1.28	1.10
		zfve	0.90	0.94	0.81	0.74	0.54	0.30
	ΔSSIM	zfvf	0.013	0.014	0.017	0.025	0.040	0.053
		zfve	0.002	0.001	0.001	0.000	-0.004	-0.007
Pantomime	ΔPSNR	zfvf	3.40	3.46	3.09	2.87	2.61	2.08
		zfve	1.13	1.24	1.17	1.15	1.02	0.68
	ΔSSIM	zfvf	0.002	0.002	0.003	0.006	0.010	0.019
		zfve	0.000	0.000	0.000	0.000	0.000	-0.001
Champagne	ΔPSNR	zfvf	1.16	0.97	1.04	0.84	0.68	0.78
		zfve	1.14	0.95	1.12	0.94	0.88	0.58
	ΔSSIM	zfvf	0.000	0.000	0.003	0.004	0.005	0.023
		zfve	0.004	0.003	0.003	0.003	0.006	-0.003
Dog	ΔPSNR	zfvf	0.82	0.64	0.41	0.23	0.10	0.03
		zfve	0.28	0.20	0.15	0.09	0.10	0.08
	ΔSSIM	zfvf	0.006	0.006	0.006	0.006	0.007	0.014
		zfve	0.000	0.000	0.000	0.000	-0.001	-0.005
Kendo	ΔPSNR	zfvf	0.81	0.69	0.69	0.87	1.03	1.16
		zfve	1.00	1.20	1.28	0.89	0.61	0.53
	ΔSSIM	zfvf	0.002	0.003	0.007	0.012	0.022	0.033
		zfve	0.002	0.002	0.002	0.001	0.000	0.002

with respect to *Lanczos* method are 1.39 and −0.16 dB, respectively. While the average PSNR gains of the proposed method in comparison with *Lanczos* method at the same QP values are 3.21 and 0.74 dB, respectively.

The subjective comparisons are shown in Fig. 8. The reference FR frame at $QP = 32$ is shown in Fig. 8(a), whereas Fig. 8(b) shows a cropped portion of it, the same areas processed by the benchmark and the proposed method are shown in Fig. 8(c) and (d), respectively. In contrast to *Lanczos*, our method preserves the edges and obtains a satisfactory result, due to the elimination of the aliasing artifacts and blurring caused by only adopting interpolation process.

B. Performance of Each Stage of the Proposed Method

In this section, several experiments have been conducted to validate the necessity and effectiveness of each stage in the proposed algorithm. Hence, the PSNR and SSIM improvements of each stage are listed in Table V for the zero-filled view filling stage and the enhancement stage, these two stages will be, respectively, denoted as zfvf and zfve for short. As can be seen from the table, each stage contributes some gains, except for some very small losses. For the majority of the tested sequences, the first stage provides significant gains, nevertheless, the second stage also makes some smaller contributions by doing some local improvement, when QP is large. Similarly, the first stage provides more SSIM gains than the second stage, except for the sequence *Champagne*.

To further investigate the effectiveness of each stage the sequence *Doorflower* is taken as an example in Fig. 9. The PSNR value for each 8 × 8 block has been shown for the *Lanczos* method in Fig. 9(a) [the original tested frame is shown in Fig. 8(a)]. It is worth noting that it has over 50 dB PSNR values at the top-right corner of the frame and also high PSNR values at smooth areas. However, areas with complex texture and edges suffer low PSNR values, some of these areas are indicated by red squares in Fig. 9. This observation emphases the importance of recovering high-frequency information and the weakness of approaches that solely rely on the LR pixels to generate the FR frame. By referring to Fig. 9(b), which shows the PSNR distribution after replacing the zero-filled pixels in the zero-filled view, we could observe that some parts (especially in the red squares highlighted parts) the edges and areas with complex texture

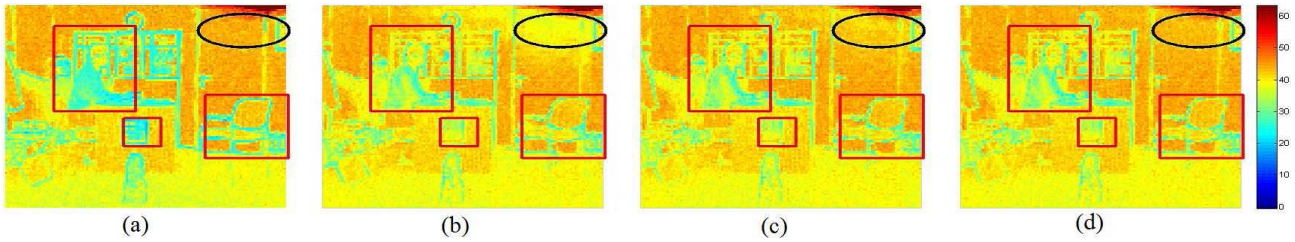


Fig. 9. PSNR value for each 8 × 8 block evaluated on the luminance component of the first frame of *Doorflower* [as shown in Fig. 8(a)] at QP = 22. (a) Benchmark interpolation method. Proposed method: (b) after similarity check, (c) after smoothness check, (d) after enhancement stage.

TABLE VI
LUMINANCE PSNR (dB) AND SSIM VALUES AND GAINS OVER THE BENCHMARK METHOD FOR MULTIVIEW VIDEO

Doorflower							
QP		22	27	32	37	42	47
PSNR	Lan	33.40	33.15	32.55	31.46	29.62	27.43
	Pro	**37.48**	**36.89**	**35.62**	**33.80**	**31.49**	**28.70**
ΔPSNR		4.08	3.74	3.07	2.34	1.87	1.27
SSIM	Lan	0.972	0.967	0.957	0.941	0.904	0.840
	Pro	**0.988**	**0.982**	**0.973**	**0.960**	**0.932**	**0.876**
ΔSSIM		0.015	0.015	0.016	0.019	0.028	0.036
Pantomime							
QP		22	27	32	37	42	47
PSNR	Lan	35.56	35.27	34.53	33.13	30.80	27.54
	Pro	**40.16**	**39.71**	**38.11**	**35.41**	**31.91**	**28.00**
ΔPSNR		4.60	4.44	3.58	2.28	1.11	0.46
SSIM	Lan	0.994	0.993	0.991	0.986	0.975	0.939
	Pro	**0.997**	**0.996**	**0.994**	**0.989**	**0.978**	**0.942**
ΔSSIM		0.003	0.003	0.003	0.003	0.003	0.003
Champagne							
QP		22	27	32	37	42	47
PSNR	Lan	34.01	33.67	32.90	31.62	29.55	26.95
	Pro	**37.75**	**36.88**	**35.48**	**33.17**	**30.42**	**27.31**
ΔPSNR		3.74	3.21	2.58	1.55	0.87	0.36
SSIM	Lan	0.992	0.990	0.985	0.977	0.961	0.923
	Pro	**0.996**	**0.994**	**0.989**	**0.981**	**0.965**	**0.926**
ΔSSIM		0.004	0.004	0.004	0.004	0.004	0.003
Dog							
QP		22	27	32	37	42	47
PSNR	Lan	34.99	34.29	32.95	31.07	28.45	25.82
	Pro	**36.81**	**35.77**	**33.97**	**31.71**	**28.76**	**26.00**
ΔPSNR		1.82	1.48	1.02	0.64	0.31	0.18
SSIM	Lan	0.983	0.977	0.962	0.933	0.862	0.733
	Pro	**0.991**	**0.984**	**0.970**	**0.940**	**0.869**	**0.743**
ΔSSIM		0.008	0.007	0.007	0.007	0.007	0.010
Kendo							
QP		22	27	32	37	42	47
PSNR	Lan	37.56	36.85	35.48	33.49	30.87	27.86
	Pro	**41.05**	**39.90**	**38.01**	**35.58**	**32.58**	**29.35**
ΔPSNR		3.49	3.05	2.53	2.09	1.71	1.49
SSIM	Lan	0.986	0.981	0.972	0.956	0.927	0.877
	Pro	**0.990**	**0.986**	**0.980**	**0.968**	**0.946**	**0.907**
ΔSSIM		0.004	0.005	0.008	0.012	0.020	0.030

have been improved significantly. However, at the same time the black ellipse highlighted part, which is a flat background area in the scene, endures some quality degradation due to this process. This indicates that if most of the pixels are copied from virtual view, the information of edges and some details can be recovered well, but the flat area might be degraded with respect to the interpolated frame. Hence, there is an overall tradeoff between these two choices. The FR frame after smoothness check is shown in Fig. 9(c), from this figure we could appreciate how the quality of flat areas is improved [the ellipse highlighted part in Fig. 9(c)]. Actually, the ellipse highlighted part in the scene is closer to the light source, so it has higher possibility to be affected by the imbalanced light distribution. Therefore, it gets improved after the luminance compensation process, shown in Fig. 9(d), and the already improved parts in previous stage are still preserved well.

C. Performance Evaluation on Multiview Video

When testing on multiview video, for *Doorflower*, the LR version of View10 is super-resolved with the aid of View12 and View08. For *Pantomime*, the LR version of View40 is super-resolved with the aid of View39 and View41. For *Champagne*, the LR version of View38 is super-resolved with the aid of View37 and View39. For *Dog*, the LR version of View39 is super-resolved with the aid of View38 and View40. For *Kendo*, the LR version of View04 is super-resolved with the aid of View03 and View05. Table VI reveals all the results of these simulations and it shows that the proposed SR method can also work well in multiview video system. In this case, the highest PSNR gain can be up to 4.6 dB for *Pantomime* sequence. Compared with the two-view video case, the PSNR gains of multiview video become higher, especially when the QP is small (QP = 22), and the average gain over the tested sequences for all QPs is 2.16 dB which is 0.20 dB higher than the obtained results for stereoscopic video. These gains are obtained due to the availability of multiple virtual-view candidates, which ensures that the more suitable virtual-view pixels are copied into the zero-filled view.

VI. CONCLUSION

In this paper, a novel interpolation-based virtual-view-assisted SR method for MR multiview-plus-depth video has been proposed. The LR views in the MR multiview video are super-resolved to FR size using two stages. In the first stage, the similarity between the LR pixels and their counterparts in the virtual view will be measured. Then if necessary, smoothness check will be carried out to determine whether using virtual-view pixels or interpolated pixels to

fill the zero-filled pixels. Subsequently, the quality of the virtual-view-based pixels is enhanced by compensating the intrinsic luminance difference between the two views. Furthermore, the inter-view correlation is exploited to enhance the LR pixels in the super-resolved frame by reducing their compression distortion. Therefore, different from the previous interpolation-based SR algorithms, the advantages of virtual views have been exploited by the proposed method at different stages. Moreover, it has been shown that the proposed algorithm achieves superior performance with respect to other approaches. Future work will be devoted to combine temporal with inter-view correlation to improve the exploitation of the virtual views.

APPENDIX
DERIVATION OF T_{si} AND T_l

This appendix explains the process of deriving thresholds T_{si} and T_l. The used approach stems from the idea that the LR and FR frames are affected by compression distortion, and consequently the thresholds should take this distortion into account. To derive the threshold T_{si} which is used to qualitatively indicate the local texture similarity, let us suppose that a point is projected into FR viewpoint and LR viewpoint as $\hat{V}_1^F(\mu,\nu)$ and $\hat{V}_2^L(i,j)$, respectively. As in Section II-B, let us suppose that the previous two values could be regarded similar $\hat{V}_1^F(\mu,\nu)=\hat{V}_2^L(i,j)=\hat{v}$, which means the original point is not occluded in any of the two viewing points. Then after compression $\hat{V}_1^F(\mu,\nu)$ and $\hat{V}_2^L(i,j)$ will become: $V_1^F(\mu,\nu)=\hat{v}+d_1$ and $V_2^L(i,j)=\hat{v}+d_2$, where d_1 and d_2 are distortions caused by video compression on views 1 and 2, respectively. The mean and variance of d_1 and d_2 are $E\{d_1\}=0$, $\sigma_1^2=E\{d_1^2\}$ and $E\{d_2\}=0$, $\sigma_2^2=E\{d_2^2\}$, respectively. At the decoder side, the zero-filled view is obtained from the LR view by inserting zeros in between its pixels, thus

$$V_2^Z(2i-1,2j-1)=V_2^L(i,j)=\hat{v}+d_2. \tag{12}$$

Assuming that the wrapping process works accurately and maps $V_1^F(\mu,\nu)$ into $V_{1,2}^V(2i-1,2j-1)$ without introducing tangible wrapping distortion. In this case

$$V_{1,2}^V(2i-1,2j-1)=\hat{v}+d_1. \tag{13}$$

Therefore, at this stage the variance of the difference between $V_{1,2}^V$ and V_2^Z, which will be used to measure the local texture similarity, could be evaluated as

$$\begin{aligned}\sigma_d^2 &= E\left\{\left(V_{1,2}^V(2i-1,2j-1)-V_2^Z(2i-1,2j-1)\right)^2\right\}\\ &= E\{(d_1-d_2)^2\}.\end{aligned} \tag{14}$$

Due to the fact that d_1 and d_2 are uncorrelated, $E\{(d_1-d_2)^2\}=E\{\sigma_1^2\}+E\{\sigma_2^2\}$. So when measuring the local similarity, the threshold T_{si} should be selected to mask the distortion induced dissimilarity, thus

$$T_{\mathrm{si}}=\alpha\times\sigma_d=\alpha\sqrt{\sigma_1^2+\sigma_2^2} \tag{15}$$

where α is a parameter which depends on the sequence content.

The derivation of the luminance compensation threshold, T_l, follows a similar approach to the one used for T_{si}. The luminance compensation process is carried forward for the virtual-view-based recovered pixels. This to happen requires that the similarity condition between W^Z and W^V be satisfied. Thus, we could use the same approach we used to evaluate σ_d^2 in (14) to evaluate the variance of ΔY^h as

$$\sigma_h^2=\frac{\sigma_1^2+\sigma_2^2}{2}. \tag{16}$$

For the vertical compensation item ΔY^v the variance could be evaluated as

$$\sigma_v^2=\frac{\sigma_1^2+\sigma_2^2}{2}. \tag{17}$$

Finally, for the center compensation item ΔY^c we have

$$\sigma_c^2=\frac{\sigma_1^2+\sigma_2^2}{4}. \tag{18}$$

Hence, if we want to use threshold T_l to ensure that mainly luminance differences get compensated and not the differences due to compression of the two views, then T_l should be selected to be larger than σ_h, σ_v, and σ_c.

Therefore

$$T_l=\beta\sqrt{\sigma_1^2+\sigma_2^2} \tag{19}$$

where β is a factor that depends on the sequence.

ACKNOWLEDGMENT

The authors would like to thank Prof. Z. Wang's Ph.D. student, J. Zhang, with the Department of Automation, University of Science and Technology of China, for the support in the technical aspects.

REFERENCES

[1] C. Fehn, K. Hopf, and B. Quante, "Key technologies for an advanced 3D TV system," *Proc. SPIE*, vol. 5599, pp. 66–80, Oct. 2004.
[2] M. Tanimoto, "Overview of free viewpoint television," *Signal Process., Image Commun.*, vol. 21, no. 6, pp. 454–461, Mar. 2006.
[3] J. Xiao, M. M. Hannuksela, T. Tillo, M. Gabbouj, C. Zhu, and Y. Zhao, "Scalable bit allocation between texture and depth views for 3-D video streaming over heterogeneous networks," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 1, pp. 139–152, Jan. 2015.
[4] H. S. Sawhney, Y. Guo, K. Hanna, R. Kumar, S. Adkins, and S. Zhou, "Hybrid stereo camera: An IBR approach for synthesis of very high resolution stereoscopic image sequences," in *Proc. 28th Annu. Conf. SIGGRAPH*, 2001, pp. 451–460.
[5] P. Aflaki, M. M. Hannuksela, J. Hakkinen, P. Lindroos, and M. Gabbouj, "Subjective study on compressed asymmetric stereoscopic video," in *Proc. 17th IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2010, pp. 4021–4024.
[6] X. Li, Y. Hu, X. Gao, D. Tao, and B. Ning, "A multi-frame image super-resolution method," *Signal Process.*, vol. 90, no. 2, pp. 405–414, Feb. 2010.
[7] X. Gao, Q. Wang, X. Li, D. Tao, and K. Zhang, "Zernike-moment-based image super resolution," *IEEE Trans. Image Process.*, vol. 20, no. 10, pp. 2738–2747, Oct. 2011.
[8] W. T. Freeman, T. R. Jones, and E. C. Pasztor, "Example-based super-resolution," *IEEE Comput. Graph. Appl.*, vol. 22, no. 2, pp. 56–65, Mar./Apr. 2002.
[9] X. Gao, K. Zhang, D. Tao, and X. Li, "Joint learning for single-image super-resolution via a coupled constraint," *IEEE Trans. Image Process.*, vol. 21, no. 2, pp. 469–480, Feb. 2012.
[10] L. Zhang and X. Wu, "An edge-guided image interpolation algorithm via directional filtering and data fusion," *IEEE Trans. Image Process.*, vol. 15, no. 8, pp. 2226–2238, Aug. 2006.

[11] M. Li and T. Q. Nguyen, "Markov random field model-based edge-directed image interpolation," *IEEE Trans. Image Process.*, vol. 17, no. 7, pp. 1121–1128, Jul. 2008.
[12] K. Zhang, X. Gao, D. Tao, and X. Li, "Single image super-resolution with non-local means and steering kernel regression," *IEEE Trans. Image Process.*, vol. 21, no. 11, pp. 4544–4556, Nov. 2012.
[13] H. Su, L. Tang, Y. Wu, D. Tretter, and J. Zhou, "Spatially adaptive block-based super-resolution," *IEEE Trans. Image Process.*, vol. 21, no. 3, pp. 1031–1045, Mar. 2012.
[14] D. C. Garcia, C. Dorea, and R. L. de Queiroz, "Super resolution for multiview images using depth information," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 9, pp. 1249–1256, Sep. 2012.
[15] J. Zhang, Y. Cao, and Z. Wang, "A simultaneous method for 3D video super-resolution and high-quality depth estimation," in *Proc. 20th IEEE ICIP*, Sep. 2013, pp. 1346–1350.
[16] A. K. Jain and T. Q. Nguyen, "Video super-resolution for mixed resolution stereo," in *Proc. 20th IEEE ICIP*, Sep. 2013, pp. 962–966.
[17] P. Aflaki, M. M. Hannuksela, J. Hakkinen, P. Lindroos, and M. Gabbouj, "Impact of downsampling ratio in mixed-resolution stereoscopic video," in *Proc. 3DTV-Conf., True Vis.-Capture, Transmiss., Display 3D Video (3DTV-CON)*, Jun. 2010, pp. 1–4.
[18] L. Stelmach, W. J. Tam, D. Meegan, and A. Vincent, "Stereo image quality: Effects of mixed spatio-temporal resolution," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 10, no. 2, pp. 188–193, Mar. 2000.
[19] L. McMillan, Jr., "An image-based approach to three-dimensional computer graphics," Ph.D. dissertation, Dept. Comput. Sci., Univ. North Carolina Chapel Hill, Chapel Hill, NC, USA, 1997.
[20] B. Shi, Y. Li, L. Liu, and C. Xu, "Color correction and compression for multi-view video using H.264 features," in *Computer Vision*. Berlin, Germany: Springer-Verlag, 2010, pp. 43–52.
[21] J. Lévy-Véhel and P. Legrand, "Hölderian regularity-based image interpolation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, vol. 3. May 2006, pp. 852–855.
[22] A. M. Tourapis, K. Sühring, and G. Sullivan, *H.264/14496-10 AVC Reference Software Manual*, document JVT-AE010 ISO/IEC JTC1/SC29/WG11 and ITU-T SG16 Q.6, London, U.K., 2009.
[23] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.
[24] J. Yang, J. Wright, T. S. Huang, and Y. Ma, "Image super-resolution via sparse representation," *IEEE Trans. Image Process.*, vol. 19, no. 11, pp. 2861–2873, Nov. 2010.

Zhi Jin (S'14) received the B.S. degree in telecommunication engineering from University of Liverpool, Liverpool, U.K., and Xi'an Jiaotong-Liverpool University, Suzhou, China, in 2011, where she is currently working toward the Ph.D. degree.

Her research interests include video compression and processing, 3-D video coding, and applications of depth maps.

Tammam Tillo (M'05–SM'12) received the Engineer Diploma degree in electrical engineering from University of Damascus, Damascus, Syria, in 1994 and the Ph.D. degree in electronics and communication engineering from Politecnico di Torino, Turin, Italy, in 2005.

He was a Visiting Researcher with École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland, in 2004. He was a Post-Doctoral Researcher with the Image Processing Laboratory, Politecnico di Torino, from 2005 to 2008. He was with Xi'an Jiaotong-Liverpool University (XJTLU), Suzhou, China, in 2008. He was an Invited Research Professor with the Digital Media Laboratory, SungKyunKwan University, Seoul, Korea. He was promoted to Full Professor in 2012. He was the Head of the Department of Electrical and Electronic Engineering at XJTLU from 2010 to 2013, and the Acting Head of the Department of Computer Science and Software Engineering from 2012 to 2013. He is an Expert Evaluator of several national-level research programs. His research interests include robust transmission of multimedia data, image and video compression, and hyperspectral image compression.

Chao Yao received the B.S. degree in computer science from Beijing Jiaotong University, Beijing, China, in 2009, where he is currently working toward the Ph.D. degree in signal and information processing with the Institute of Information Science.

His research interests include video compression and processing, 3-D video coding, and 3-D computer vision.

Jimin Xiao (M'14) received the B.S. and M.E. degrees in telecommunication engineering from Nanjing University of Posts and Telecommunications, Nanjing, China, in 2004 and 2007, respectively, and the Ph.D. degree in electrical engineering and electronics from University of Liverpool, Liverpool, U.K., in 2013.

He was a Visiting Researcher with Nanyang Technological University, Singapore, in 2013, and a Senior Researcher with the Department of Signal Processing, Tampere University of Technology, Tampere, Finland, and an External Researcher with the Nokia Research Center, Tampere, from 2013 to 2014. He has been with Xi'an Jiaotong-Liverpool University, Suzhou, China, since 2014, as a Faculty Member. His research interests include video streaming, image and video compression, and multiview video coding.

Yao Zhao (M'06–SM'12) received the B.S. degree from the Department of Radio Engineering, Fuzhou University, Fuzhou, China, in 1989; the M.E. degree from the Department of Radio Engineering, Southeast University, Nanjing, China, in 1992, and the Ph.D. degree from the Institute of Information Science, Beijing Jiaotong University (BJTU), Beijing, China, in 1996.

He became an Associate Professor with BJTU in 1998 and became a Professor in 2001. He was a Senior Research Fellow with the Information and Communication Theory Group, Faculty of Information Technology and Systems, Delft University of Technology, Delft, The Netherlands, from 2001 to 2002. He is currently the Director of the Institute of Information Science at BJTU. He is currently leading several national research projects from the 973 Program, 863 Program, and the National Science Foundation of China. His research interests include image/video coding, digital watermarking and forensics, and video analysis and understanding.

He is on the Editorial Boards of several international journals, including as an Associate Editor of IEEE TRANSACTIONS ON CYBERNETICS, IEEE SIGNAL PROCESSING LETTERS, and *Circuits, System, and Signal Processing* (Springer), and an Area Editor of *Signal Processing: Image Communication* (Elsevier). He was named as a Distinguished Young Scholar from the National Science Foundation of China in 2010, and was elected as a Chang Jiang Scholar from the Ministry of Education, China, in 2013.