Virtual View Assisted Video Super-Resolution and Enhancement

Zhi JIN, Student Member, IEEE, Tammam TILLO, Senior Member, IEEE, Chao YAO, Jimin XIAO, Member, IEEE, Yao ZHAO, Senior Member, IEEE

Abstract—3D multiview video provides users an experience different from traditional video, however, it puts huge burden on the limited bandwidth resources. Mixed-resolution video in multiview system can alleviate this problem by using different video resolutions for different views. However, in order to reduce visual uncomfortableness and to make this video format more suitable for free-viewpoint television, the low resolution 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 to fill the missing pixels in the super-resolved frame by virtual view pixels or by spatially 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 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 low resolution 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 PSNR gain of up to 3.85dB.

Index Terms—Virtual view, Interpolation-based Super-Resolution, Mixed resolution system, Multiview video

I. INTRODUCTION

With the development of video technology, 3D video applications are increasingly accessible to customers. Among these applications, Three-Dimensional Television (3DTV) [1] and Free-viewpoint Television (FTV) [2] have attracted a lot of attention. 3DTV provides viewers the perception of real-world scenes by 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 3D video relies on 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] [5] that uses Mixed Resolution (MR) video, in which at least one of the views is captured at Low Resolution (LR), while the others are captured at 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, in order to meet the requirement 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 by using the Super Resolution (SR) techniques at the decoder side [4]. Therefore, in a 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: multi-image-based SR algorithms [6] [7], example-learning-based SR algorithms [8] [9] and 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 sub-pixel misalignments. These approaches can be realized on both frequency domain and spatial domain [6] [7]. However, this kind of methods highly rely 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 by 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 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). In order to overcome some of these weaknesses, Zhang et.al proposed to adaptively fuse the LR pixels on the two diagonal directions according to the Linear Minimum Mean Squares-error Estimation (LMMSE) technique [10]. In [11], Li et.al 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. in [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 3D 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 3D MRF model was used to find the optimized patches from a database of HR images both in spatial and temporal domain 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, firstly, the high frequency information contained in FR views can be properly utilized to super-resolve LR views; secondly, 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 experimental results are presented in Section V. Section VI concludes the 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 resolution factor two by given the findings in [17] and [18], and the framework of the proposed virtual view assisted interpolation-based SR algorithm is depicted 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 \( V_I^F \) and \( D_I^F \), respectively, while, the decoded LR texture sequence will be denoted by \( V_Z^F \). At the decoder side, the decompressed LR view is used to generate two intermediate FR versions at viewpoint 2. The first version \( V_{Z,1}^F \) is obtained by using an interpolation method, such as bilinear or bicubic. The second version \( V_{Z,2}^F \) 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.

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 in \( V_Z^F \) 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 \( V_Z^F \). This process will minimize the probability of copying some holes from the virtual view into \( V_Z^F \). The similarity check mechanism consists of two \( 3 \times 3 \) 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 \( W_Z^V(-1, -1) \) stands for the upper left corner pixel in the window \( W_Z^V \). 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 \(1 \leq i \leq H/2\) and \(1 \leq j \leq W/2\). This ensures that for the zero-filled view, there are four LR pixels at the corners of the window \( W_Z^V \) to measure the local texture similarity between the zero-filled view and the virtual view.

### A. Zero-filled View Filling

![A pictorial representation of the similarity check process and the generation of FR frame.](image-url)
In this paper, the Sum of Absolute Difference (SAD) is used for this purpose as below:\footnote{The SAD and Euclidean distance in this case will almost lead to the same results.}

\[
D_{SC} = \sum_{\eta \in \{-1,1\}, \theta \in \{-1,1\}} |W^Z(\eta, \theta) - W^V(\eta, \theta)| \tag{1}
\]

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 \(V^I_1\) if the SAD value is larger than a threshold \(T_{\text{si}}\), as shown in the following equation:

\[
V^Z_2(\eta, \theta) = V^I_1(\eta, \theta) ; D_{SC} \geq T_{\text{si}} \tag{2}
\]

where \((\eta, \theta) \in C\) and \(C = \{(2i, 2j), (2i - 1, 2j), (2i, 2j - 1)\}\). 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 \(T_{\text{si}}\), 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 \(V^Z_2\). 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\footnote{The SAD and Euclidean distance in this case will almost lead to the same results.}. Hence, based on this fact zero-filled pixels in smooth areas will be replaced by their counterparts from \(V^Z_1\). On the other hand, for non-smooth 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 non-smooth areas. Thus, for this kind of areas the zero-filled pixels will be replaced by their counterparts in the virtual view \(V^1_2\).

The previous paradigm is implemented in the second step, where the smoothness of a \(3 \times 3\) area, \(W^I\), centered at the pixel \((2i, 2j)\) in \(V^Z_2\) is checked; in this paper, this has been done by measuring the standard deviation, \(\sigma_s\). The motivation behind using the window \(W^I\) to measure the local smoothness, is that a non-trivial interpolator uses more than 8-connected neighbors in the estimation process to preserve the local regularity\footnote{The SAD and Euclidean distance in this case will almost lead to the same results.}. 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 the following equation:

\[
V^Z_2(\eta, \theta) = \begin{cases} V^I_1(\eta, \theta) & ; \sigma_s < T_{\text{sym}} \\ V^Z_1(\eta, \theta) & ; \sigma_s \geq T_{\text{sym}} \end{cases} \tag{3}
\]

where \((\eta, \theta) \in C\). In Eq.(3), \(T_{\text{sym}}\) is a threshold to determine whether an area surrounding the pixel \((2i, 2j)\) has smooth or non-smooth 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.

\begin{figure}[h]
\centering
\includegraphics[width=\columnwidth]{fig1.png}
\caption{The framework of the proposed super-resolution method.}
\end{figure}

\begin{figure}[h]
\centering
\includegraphics[width=\columnwidth]{fig3.png}
\caption{Flowchart of the Zero-filled View Filling stage.}
\end{figure}

\section{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.

**Table I**
The PSNR differences between FR and LR views using H.264 for: (A) bookarrival; (B) doorflower; (C) laptop; (D) champagne

<table>
<thead>
<tr>
<th>Seq</th>
<th>Bookarrival</th>
<th>Doorflower</th>
<th>Laptop</th>
<th>Champagne</th>
</tr>
</thead>
<tbody>
<tr>
<td>QP</td>
<td>PSNR(dB)</td>
<td>PSNR(dB)</td>
<td>PSNR(dB)</td>
<td>PSNR(dB)</td>
</tr>
<tr>
<td>22</td>
<td>1.00</td>
<td>0.78</td>
<td>0.83</td>
<td>0.65</td>
</tr>
<tr>
<td>27</td>
<td>1.48</td>
<td>1.15</td>
<td>1.22</td>
<td>1.28</td>
</tr>
<tr>
<td>32</td>
<td>2.07</td>
<td>1.88</td>
<td>1.82</td>
<td>2.23</td>
</tr>
<tr>
<td>37</td>
<td>2.55</td>
<td>2.42</td>
<td>2.31</td>
<td>2.75</td>
</tr>
<tr>
<td>42</td>
<td>2.69</td>
<td>2.68</td>
<td>2.56</td>
<td>2.80</td>
</tr>
<tr>
<td>47</td>
<td>2.22</td>
<td>2.20</td>
<td>2.19</td>
<td>3.05</td>
</tr>
</tbody>
</table>

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 $V^2_F$, 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 $V^2_F$ (i.e., the virtual-view-based recovered pixels) to have harmonious brightness with the surrounding LR pixels. If the set of pixels $C$ 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:

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

If the absolute value of $D_{LC}$ is larger than a threshold $T_i$, the compensation process will be used to update the intensity of the pixels $C$. The reason behind using the threshold $T_i$ is to eliminate the effect of compression distortion on the luminance compensation process. In fact, given that a 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 $C$, then for each of its three pixels a proper amount of compensation will be determined by using the closest available neighbors for each of the pixel in $C$ 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 $\Delta Y^h$:

$$\Delta Y^h = \frac{1}{2} [W^Z(\eta, \theta) - W^V(\eta, \theta)]$$

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 $V^2_F(2i - 1, 2j) + \Delta Y^h$ instead of $V^1_F(2i - 1, 2j)$. Similarly, the compensation value for the pixel $(2i, 2j - 1)$ will be computed starting from its two vertical neighbors as

$$\Delta Y^v = \frac{1}{2} [W^Z(\eta, \theta) - W^V(\eta, \theta)]$$

and consequently $V^2_F(2i, 2j - 1) = V^1_F(2i, 2j - 1) + \Delta Y^v$.

As for the center pixel, $(2i, 2j)$, it will be updated as $V^2_F(2i, 2j) = V^1_F(2i, 2j) + \Delta Y^c$. However, given that the pixel $(2i, 2j)$ is at equal distance from the four corners its compensation value will be evaluated as:

$$\Delta Y^c = \frac{1}{4} \sum_{\eta \in \{-1, 1\}, \theta \in \{-1, 1\}} [W^Z(\eta, \theta) - W^V(\eta, \theta)]$$

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

2) **LR pixels enhancement in the super-resolved view:** Previous subsection proposes an enhancement mechanism for the virtual-view-based recovered pixels, whereas, in this subsection a mechanism to enhance the quality of the other pixels in $V^2_F$, 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 subsection 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 viewpoint 1 and viewpoint 2 by $V^1_F(\mu, \nu)$ and $V^2_F(i,j)$, 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., $V^1_F(\mu, \nu) \approx V^2_F(i,j) = \hat v$, the smaller the baseline is the more correct this assumption is. Hence, in the following sections, we will assume that $V^1_F(\mu, \nu) = V^2_F(i,j) = \hat v$. These two projections will be compressed separately in the two viewpoints, so they become: $V^1_F(\mu, \nu) = \hat v + d_1$ and $V^2_F(i,j) = \hat v + d_2$, where $d_1$ and $d_2$ are the distortion caused by video compression on view 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 view 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

---

1Although the coordinate system in the two views are related, however, they are different due to the fact that the two views have different resolutions.
The article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2015.2412791, IEEE Transactions on Circuits and Systems for Video Technology

is obtained from the LR view by inserting zeros in between its pixels, thus:

\[ V_{i,j}^F(2i-1,2j-1) = V_{i,j}^L = \hat{v} + d_2 \]  

(7)

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

\[ V_{1/2}(2i-1,2j-1) = \hat{v} + d_1 \]  

(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 \( V_{i,j}^F(2i-1,2j-1) \) and \( V_{1/2}(2i-1,2j-1) \) then the expected compression distortion could be evaluated by using (12) and (13) as:

\[ \sigma^2 = E \left\{ \left( \frac{v_d + d_2 + \hat{v} + d_1 - \hat{v}}{2} \right)^2 \right\} \]  

(9)

Since view 1 and 2 are separately compressed, \( E\{d_1d_2\} = 0 \). In the general case \( d_1 \leq d_2 \) even when using the same QP for the LR and FR views, consequently, \( \frac{\sigma^2}{4} \leq \sigma^2 \leq \frac{\sigma^2}{2} \). This means that the equivalent distortion of the pixels at \((2i-1,2j-1)\), where \( 1 \leq i \leq H/2 \) and \( 1 \leq j \leq W/2 \) in the zero-filled view will be reduced.

It is worth noticing that the averaging process can only be applied to those pixels in \( V_{i,j}^F \) which have equivalent pixels in \( V_{1/2}^V \), 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 \( \frac{V_{i,j}^F(2i-1,2j-1) + V_{1/2}(2i-1,2j-1)}{2} \).

III. Thresholds Evaluation

In the proposed SR approach, both the virtual view and interpolated view are utilized to generate the FR frames, and two post-processing enhancement operations are exploited to further improve the quality of the generated FR view. In this whole process, three thresholds are required. An experimental-based approach to determine the values of these thresholds could be used at the encoder side by 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 nestlike loops, and then send values to the decoder. If the complexity of estimating one threshold is \( O(n) \), then the complexity of this exhaustive approach is \( O(n^3) \). 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 \( O(n^3) \) to \( O(3n) \). Some experiments were conducted on “Doorflower” and “Pantomime” sequences.
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. 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 Appendix that $T_{si}$ and $T_l$ could be represented by

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

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

where $\sigma_1$ and $\sigma_2$ are the standard deviation of the compression distortion affecting view 1 (FR view) and view 2 (LR view), respectively. $\alpha$ and $\beta$ are two parameters which depend on the sequence content.

In the successive approach, the three thresholds (or equivalently, $\alpha$, $\beta$ 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 which uses a user-defined value for both $\alpha$ and $\beta$ and the corresponding results are also shown in Fig.6. In this approach, it is reasonable to assume $\alpha$ and $\beta$ are larger than three based on Chebyshev’s inequality. From the figure, it is obvious that the successive-search approach which estimates $\alpha$, $\beta$ 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.

IV. MULTIVIEW VIDEO

The proposed virtual view assisted SR algorithm can also be applied to multiview multi-resolution 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 depicted in Fig.7, $V_q^k$ ($q = 1, ..., m$ and $k = 1, ..., n$) 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. Consequently, the two enhancement methods are carried out step by step. In this way, the proposed algorithm effectively super-resolves the LR views.

Figure 7. The proposed algorithm for multiview multi-resolution system.

### V. EXPERIMENTAL RESULTS

**Table II**

<table>
<thead>
<tr>
<th>Name</th>
<th>Size</th>
<th>FR</th>
<th>LR</th>
<th>Content’s Motion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Doorflower</td>
<td>1024 × 768</td>
<td>View10</td>
<td>View08</td>
<td>Moderate</td>
</tr>
<tr>
<td>Bookarrival</td>
<td>1024 × 768</td>
<td>View08</td>
<td>View06</td>
<td>Moderate</td>
</tr>
<tr>
<td>Leavelaptop</td>
<td>1024 × 768</td>
<td>View06</td>
<td>View07</td>
<td>Moderate</td>
</tr>
<tr>
<td>Champagne</td>
<td>1280 × 960</td>
<td>View39</td>
<td>View40</td>
<td>Medium complex</td>
</tr>
<tr>
<td>Dog</td>
<td>1280 × 960</td>
<td>View37</td>
<td>View38</td>
<td>Complex</td>
</tr>
<tr>
<td>Kendo</td>
<td>1024 × 768</td>
<td>View03</td>
<td>View04</td>
<td>Complex</td>
</tr>
</tbody>
</table>

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

In order 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 downsampling at least one of the FR views to LR for each of the tested sequences. The downsampling factor is 2 on both horizontal and vertical direction. 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 6-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.
Figure 6. PSNR and SSIM comparisons of different approaches for the evaluation of α and β; a) and b) are results of “Doorflower”; c) and d) are results of “Pantomime”.

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”, “Pro”= “Proposed”. Results of the state-of-the-art single image super-resolution approach via Sparse Coding (SC) [24] are also reported in the table, where the parameters for the publicly available code4 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 ∆PSNR and ∆SSIM, 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 3.85dB on the sequence “Bookarrival” when PQ=22, while, the average PSNR gain over all sequences and QPs is 2.11dB. 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.

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Book</th>
<th>Doorflower</th>
<th>Laptop</th>
</tr>
</thead>
<tbody>
<tr>
<td>PSNR</td>
<td>33.04</td>
<td>33.27</td>
<td>33.95</td>
</tr>
<tr>
<td>∆PSNR</td>
<td>1.53</td>
<td>0.49</td>
<td>0.45</td>
</tr>
<tr>
<td>SSIM</td>
<td>0.941</td>
<td>0.941</td>
<td>0.945</td>
</tr>
<tr>
<td>∆SSIM</td>
<td>0.045</td>
<td>0.042</td>
<td>0.041</td>
</tr>
</tbody>
</table>

In the following we compare our proposed approach with [14]. The proposed method was tested under disadvantageous condition with respect to [14], where in the later approach the uncompressed sequence “Pantomime” and “Dog” with
<table>
<thead>
<tr>
<th>QP</th>
<th>Sequence</th>
<th>Discrete Wavelet</th>
<th>Biorthogonal</th>
<th>LL (deaplay)</th>
<th>Panorama</th>
<th>Changemaker</th>
<th>Dog</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>22</td>
<td>PSNR</td>
<td>15.12 32.09</td>
<td>33.31</td>
<td>19.50</td>
<td>33.95</td>
<td>34.90</td>
<td>37.56</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>15.26 33.07</td>
<td>35.40</td>
<td>18.54</td>
<td>34.60</td>
<td>35.50</td>
<td>37.40</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ΔPSNR</td>
<td>0.070 0.970</td>
<td>0.970 0.984</td>
<td>0.992 0.083</td>
<td>0.088</td>
<td>0.046</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>SSIM</td>
<td>0.070 0.970</td>
<td>0.970 0.984</td>
<td>0.992 0.083</td>
<td>0.088</td>
<td>0.046</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Pro.</td>
<td>0.087 0.087</td>
<td>0.087 0.087</td>
<td>0.095 0.095</td>
<td>0.095</td>
<td>0.095</td>
<td></td>
<td></td>
</tr>
<tr>
<td>37</td>
<td>PSNR</td>
<td>31.93 30.89</td>
<td>34.80</td>
<td>21.31</td>
<td>32.62</td>
<td>34.22</td>
<td>36.65</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>32.52 32.73</td>
<td>33.09</td>
<td>18.25</td>
<td>33.67</td>
<td>34.35</td>
<td>36.49</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ΔPSNR</td>
<td>0.017 0.017</td>
<td>0.017 0.017</td>
<td>0.017 0.017</td>
<td>0.017</td>
<td>0.017</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>SSIM</td>
<td>0.017 0.017</td>
<td>0.017 0.017</td>
<td>0.017 0.017</td>
<td>0.017</td>
<td>0.017</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Pro.</td>
<td>0.017 0.017</td>
<td>0.017 0.017</td>
<td>0.017 0.017</td>
<td>0.017</td>
<td>0.017</td>
<td></td>
<td></td>
</tr>
<tr>
<td>42</td>
<td>PSNR</td>
<td>33.31 32.01</td>
<td>33.09</td>
<td>32.01</td>
<td>33.09</td>
<td>32.09</td>
<td>33.09</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>33.99 33.40</td>
<td>33.40</td>
<td>33.40</td>
<td>33.40</td>
<td>33.40</td>
<td>33.40</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ΔPSNR</td>
<td>0.001 0.001</td>
<td>0.001 0.001</td>
<td>0.001 0.001</td>
<td>0.001</td>
<td>0.001</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>SSIM</td>
<td>0.001 0.001</td>
<td>0.001 0.001</td>
<td>0.001 0.001</td>
<td>0.001</td>
<td>0.001</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Pro.</td>
<td>0.001 0.001</td>
<td>0.001 0.001</td>
<td>0.001 0.001</td>
<td>0.001</td>
<td>0.001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>47</td>
<td>PSNR</td>
<td>33.93 33.09</td>
<td>33.93</td>
<td>33.09</td>
<td>33.93</td>
<td>33.09</td>
<td>33.93</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>33.99 33.40</td>
<td>33.40</td>
<td>33.40</td>
<td>33.40</td>
<td>33.40</td>
<td>33.40</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ΔPSNR</td>
<td>0.001 0.001</td>
<td>0.001 0.001</td>
<td>0.001 0.001</td>
<td>0.001</td>
<td>0.001</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>SSIM</td>
<td>0.001 0.001</td>
<td>0.001 0.001</td>
<td>0.001 0.001</td>
<td>0.001</td>
<td>0.001</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Pro.</td>
<td>0.001 0.001</td>
<td>0.001 0.001</td>
<td>0.001 0.001</td>
<td>0.001</td>
<td>0.001</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
resolution factor 2 was used, and the reported gains were 2.57dB and 1.06dB over the *Lanczos* method, respectively, while, in the proposed method, even with video compression (QP=22) the gains are 3.62dB and 1.22dB, respectively. The average PSNR gain in [14] at QP = {22, 27, 32, 37} on these two sequences with respect to *Lanczos* method are 1.39dB and -0.16dB, respectively. While, the average PSNR gains of the proposed method in comparison with *Lanczos* method at the same QP values are 3.21dB and 0.74dB, 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 Fig.8 (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 subsection, 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 noticing that it has over 50dB 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 the figure. This observation emphasizes 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 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

### Table V

<table>
<thead>
<tr>
<th>QP</th>
<th>zfvf</th>
<th>zfve</th>
<th>zfvf</th>
<th>zfve</th>
</tr>
</thead>
<tbody>
<tr>
<td>22</td>
<td>3.28</td>
<td>0.016</td>
<td>3.08</td>
<td>0.017</td>
</tr>
<tr>
<td>27</td>
<td>2.58</td>
<td>0.018</td>
<td>2.16</td>
<td>0.018</td>
</tr>
<tr>
<td>32</td>
<td>2.16</td>
<td>0.023</td>
<td>1.52</td>
<td>0.024</td>
</tr>
<tr>
<td>37</td>
<td>1.52</td>
<td>0.034</td>
<td>0.95</td>
<td>0.034</td>
</tr>
<tr>
<td>42</td>
<td>0.95</td>
<td>0.034</td>
<td>0.38</td>
<td>0.044</td>
</tr>
<tr>
<td>47</td>
<td>0.38</td>
<td>0.044</td>
<td>0.11</td>
<td>0.045</td>
</tr>
</tbody>
</table>

The 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.

---

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2015.2412791, IEEE Transactions on Circuits and Systems for Video Technology
Table VI

<table>
<thead>
<tr>
<th></th>
<th>PSNR</th>
<th>PSNR</th>
<th>SSIM</th>
<th>SSIM</th>
<th></th>
<th>PSNR</th>
<th>PSNR</th>
<th>SSIM</th>
<th>SSIM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>QP=22</td>
<td>QP=27</td>
<td>QP=32</td>
<td>QP=37</td>
<td>QP=42</td>
<td>QP=47</td>
<td>QP=22</td>
<td>QP=27</td>
<td>QP=32</td>
</tr>
<tr>
<td>Dog</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Champagne</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pantomime</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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 4.6dB 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 2.16dB which is 0.20dB 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 super-resolution method for mixed-resolution multiview-plus-depth video has been proposed. The low resolution views in the MR multiview video are super-resolved to full resolution size by 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 \( V_{1}^{L}(\mu, \nu) \) and \( V_{2}^{L}(i, j) \), respectively. As we did in Section II-B, let us suppose that the previous two values could be regarded similar \( V_{1}^{L}(\mu, \nu) = V_{2}^{L}(i, j) = \hat{v} \), which means the original point is not occluded in any of the two viewing points. Then after compression \( V_{1}^{P}(\mu, \nu) \) and \( V_{2}^{P}(i, j) \) will become: \( V_{1}^{P}(\mu, \nu) = \hat{v} + d_{1} \) and \( V_{2}^{P}(i, j) = \hat{v} + d_{2} \), where \( d_{1} \) and \( d_{2} \) are distortions caused by video compression on view 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:

\[
V_{2}^{Z}(2i - 1, 2j - 1) = V_{1}^{L}(i, j) = \hat{v} + d_{2}
\]

(12)

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

\[
V_{1}^{P, 2}(2i - 1, 2j - 1) = \hat{v} + d_{1}
\]

(13)

Therefore, at this stage the variance of the difference between \( V_{1}^{P, 2} \) and \( V_{2}^{Z} \), which will be used to measure the local texture similarity, could be evaluated as:

\[
\sigma_{d}^{2} = E\left\{ (V_{1}^{P, 2}(2i - 1, 2j - 1) - V_{2}^{Z}(2i - 1, 2j - 1))^{2}\right\}
= E\left\{ (d_{1} - d_{2})^{2}\right\}
\]

(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:

\[
T_{si} = \alpha \times \sigma_{d} = \alpha \sqrt{\sigma_{1}^{2} + \sigma_{2}^{2}}
\]

(15)

where \( \alpha \) 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 \( \sigma_{h}^{2} \) in (14) to evaluate the variance of \( \Delta Y^{h} \) as:

\[
\sigma_{h}^{2} = \frac{\sigma_{1}^{2} + \sigma_{2}^{2}}{2}
\]

(16)

For the vertical compensation item \( \Delta Y^{v} \) the variance could be evaluated as:

\[
\sigma_{v}^{2} = \frac{\sigma_{1}^{2} + \sigma_{2}^{2}}{2}
\]

(17)

Finally, for the center compensation item \( \Delta Y^{c} \) we have:

\[
\sigma_{c}^{2} = \frac{\sigma_{1}^{2} + \sigma_{2}^{2}}{4}
\]

(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 \( \sigma_{h}, \sigma_{v}, \) and \( \sigma_{c} \). Therefore:

\[
T_{l} = \beta \sqrt{\sigma_{1}^{2} + \sigma_{2}^{2}}
\]

(19)

where \( \beta \) is a factor that depends on the sequence.

**ACKNOWLEDGMENT**

Thanks for some technique support from Prof. Zengfu Wang’s Ph.D. student, Jing Zhang from the Department of Automation, University of Science and Technology of China.

**REFERENCES**


Zhi JIN received the B.S degree in Telecommunication Engineering from the University of Liverpool, UK and Xi’an Jiaotong-Liverpool University (XJTLU), China, in 2011. She is currently pursuing a Ph.D. degree of University of Liverpool, UK in XJTLU. Her current research interests include video compression and processing, 3D video coding, and the applications of depth map.

Chao YAO received the B.S. degree in computer science from Beijing Jiaotong University, Beijing, China, in 2009, where he is currently pursuing the Ph.D. degree in signal and information processing at the Institute of Information Science. His current research interests include video compression and processing, 3D video coding, and 3D computer vision.

Jimin XIAO received the B.S and M.E degrees in Telecommunication Engineering from Nanjing University of Posts and Telecommunications, Nanjing, China, in 2004, 2007, respectively. He received the Ph.D degree in Electrical Engineering and Electronics from University of Liverpool, UK, in 2013. In 2013, he served as a visiting researcher at Nanyang Technological University, Singapore. From Nov. 2013 to Nov.2014, he was a senior researcher in the Department of Signal Processing, Tampere University of Technology, Finland, and external researcher in Nokia Research Center, Tampere, Finland. Since Dec. 2014 he joined Xi’an Jiaotong-Liverpool University, Suzhou, China, as a faculty member. His research interests are in the areas of video streaming, image and video compression, and multimodal video coding.

Jimin XIAO received the B.S and M.E degrees in Telecommunication Engineering from Nanjing University of Posts and Telecommunications, Nanjing, China, in 2004, 2007, respectively. He received the Ph.D degree in Electrical Engineering and Electronics from University of Liverpool, UK, in 2013. In 2013, he served as a visiting researcher at Nanyang Technological University, Singapore. From Nov. 2013 to Nov.2014, he was a senior researcher in the Department of Signal Processing, Tampere University of Technology, Finland, and external researcher in Nokia Research Center, Tampere, Finland. Since Dec. 2014 he joined Xi’an Jiaotong-Liverpool University, Suzhou, China, as a faculty member. His research interests are in the areas of video streaming, image and video compression, and multimodal video coding.

Yao ZHAO (M’06-SM’12) received the B.S. degree from Fuzhou University, Fuzhou, China, in 1989, and the M.E. degree from Southeast University, Nanjing, China, in 1992, both from the Radio Engineering Department, and the Ph.D. degree from the Institute of Information Science, Beijing Jiaotong University (BJTU), Beijing, China, in 1996. He became an Associate Professor at BJTU in 1998 and became a Professor in 2001. From 2001 to 2002, 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. He is currently the Director of the Institute of Information Science, BJTU. His current research interests include image/video coding, digital watermarking and forensics, and video analysis and understanding. He is currently leading several national research projects from the 973 Program, 863 Program, and the National Science Foundation of China. He serves on the editorial boards of several international journals, including as an Associate Editor of the IEEE Transactions on Cybernetics, Associate Editor of the IEEE Signal Processing Letters, Area Editor of Signal Processing: Image Communication (Elsevier), and Associate Editor of Circuits, System, and Signal Processing (Springer). He was named a Distinguished Young Scholar by the National Science Foundation of China in 2010, and was elected as a Chang Jiang Scholar of Ministry of Education of China in 2013.

Tamman TILLO (M’05-SM’12) received the Engineer Diploma in electrical engineering from the University of Damascus, Syria, in 1994, and the Ph.D. degree in electronics and communication engineering from Politecnico di Torino, Italy, in 2005. In 2004, he served as a Visiting Researcher at Ecole Polytechnique Federale de Lausanne (EPFL), Switzerland, and from 2005 to 2008 he worked as a Postdoctoral Researcher at the Image Processing Lab of Politecnico di Torino. For few months he was Invited Research Professor at the Digital Media Lab, SungKyunKwan University, Republic of Korea, before joining Xi’an Jiaotong-Liverpool University (XJTLU), China, in 2008. He was promoted to full professor in 2012. From 2010 to 2013 he was the Head of the Department of Electrical and Electronic Engineering at XJTLU University, and was the Acting Head of Computer Science and Software Engineering Department from 2012 to 2013. He serves as an expert evaluator for several national-level research programs. His research interests are in the areas of robust transmission of multimedia data, image and video compression, and hyperspectral image compression.