Super Resolution for Multiview Images Using Depth Information

Diogo C. Garcia, *Student Member, IEEE,* Camilo Dorea, and Ricardo L. de Queiroz, *Senior Member, IEEE*

***Abstract*—In stereoscopic and multiview video, binocular suppression theory states that the visual subjective quality of 3-D experience is not much affected by asymmetrical blurring of the individual views. Based on these studies, mixed-resolution frameworks applied for multiview systems offer great data-size reduction without incurring in significant quality degradation in 3-D video applications. However, it is interesting to recover high-frequency content of the blurred views, to reduce visual strain due to long-term exposure and to make the system suitable for free-viewpoint television. In this paper, we present a novel super-resolution technique, in which low-resolution views are enhanced with the aid of high-frequency content from neighboring full-resolution views, and the corresponding depth information for all views. Occlusions are handled by checking the consistency between views. Tests for synthetic and real image data in stereo and multiview cases are presented, and results show that significant objective quality gains can be achieved without any extra side information.**

***Index Terms*—Mixed resolution, multiview video, super resolution.**

I. Introduction

In the past decade, multiview video research has received much attention. Advances in the multiview acquisition and display technologies, the development of processors capable of higher computational capacity, and the increase of the industry interest have all contributed to a major boost in the field. Such an interest spurred several 3-D video technologies, such as home and mobile 3-DTV, free-viewpoint video, and immersive teleconferencing.

However, there are still several questions to consider. Disparity estimation between views, for instance, is still a very active research topic. Performance of such methods [1] is crucial to the success of applications ranging from coding to view rendering. Multiview compression is another significant consideration in the process, as the acquisition and transmission of several views tremendously increases the amount of data to be compressed. Roughly speaking, the data-rate increase is proportional to the number of added views, if simulcast encoding is considered. Multiview video coding (MVC) [2] can achieve a 20% rate decrease compared to simulcast encoding [3]. In this manner, a 10-view system using MVC would require a $8\times$-higher bit-rate than a single-view system's data rate, for which current networks are designed. This makes the multiview format unsuitable for many applications.

Depth perception by the human visual system presents several interesting characteristics, which might aid in reducing the bit-rate for 3-D video systems. For example, binocular suppression theory states that stereo vision is not subjectively affected by asymmetric degradation of views, depending on the circumstances [4]. Several studies have been made to determine what these circumstances are, concerning degradation by quantization and by blurring, with mixed results. Tam [5] observed that in the case of asymmetric blurring, the binocular image quality is dominated by the quality of the high-resolution view, while with asymmetric quantization, the binocular image quality corresponds to the average of both views. As for depth quality, both schemes render similar results. This indicates that a mixed-resolution (MR) coding method would be more effective in rate and subjective distortion terms. Aflaki *et al.* [6] stated that MR coding with a downsampling ratio of $1/2$ offers subjective quality similar to symmetric and quality-asymmetric full-resolution stereo coding, all under similar bit-rate constraints. MR coding would then represent a good alternative for lowering the processing complexity. Saygili *et al.* [7] indicated that for a sufficiently high quality in the reference view, users notice degradation only when the other view is encoded below a low-quality threshold that depends on the 3-D display. Furthermore, above such a threshold, users prefer asymmetric quality coding over MR coding, and below this threshold, users prefer MR coding.

These studies [4]–[7] suggest that MR coding can be a viable alternative for multiview compression, offering lower processing complexity and possibly lower bit-rates at the same subjective quality as full-resolution coding. MR stereo coding, for instance, is very suitable for mobile 3-DTV, which needs to offer a stereoscopic effect and does not afford complex view-rendering processing at the decoder side [8], [9]. Objective quality gains at low bit-rates were also reported in a MR stereo-coding framework with different downsampling ratios [10]. Temporal scalability in MR coding was also investigated [11], as well as using the high-resolution view in full resolution to predict the low-resolution view [12], eliminating the computational burden of subsampling ref-

Manuscript received March 11, 2011; revised July 21, 2011 and October 28, 2011; accepted December 30, 2011. Date of publication May 7, 2012; date of current version August 30, 2012. This work was supported in part by the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), under Grants 300223/2008-0, 470940/2010-7, and 310375/2011-8. This paper was recommended by Associate Editor M. Hannuksela.

D. C. Garcia is with the Gama Faculty and with the Department of Electrical Engineering, University of Brasilia, Brasilia 70919, Brazil (e-mail: diogo@image.unb.br).

C. Dorea and R. L. de Queiroz are with the Department of Computer Science, University of Brasilia, Brasilia 70919, Brazil (e-mail: camilo@cic.unb.br; queiroz@ieee.org).

Digital Object Identifier 10.1109/TCSVT.2012.2198134

erence frames. Image-based rendering techniques were also developed to generate a high-resolution synthesis of the low-resolution sequence at the decoder [13].

Most of these works do not attempt to estimate the high-frequency components for the low-resolution views, which can result in two major problems. First, they may not be adequate for single-view free-viewpoint television. If the user chooses a low-resolution view or an interpolated view close to the low-resolution one, quality reduction will be perceived. Second, although eye dominance does not impose a problem to MR-coding frameworks [14], [15], low-quality images continuously presented to one eye might disturb viewers in the long term [5]. Alternatively, if an intermediate low-resolution view were not available, depth-image-based rendering (DIBR) could be applied to synthesize the missing view from the adjacent full-resolution views. Nevertheless, DIBR view synthesis is susceptible to rendering artifacts, presenting objective quality drops in proportion to the distance from the adjacent views [16]. In MR setups, a low-resolution version is available, and can be used as a basis for enhancement.

In this paper, we show how the quality of low-resolution views can be objectively improved at the decoder side with the aid of the high-frequency information from neighboring views, after the low-resolution views are upsampled to full resolution. We improve on a previous work [17] by proposing a more efficient method for evaluation of interview consistency, crucial to algorithm performance. Experiments are conducted for a large set of synthetic and real images. Performance under compression is illustrated with the same test set subject to Intra coding, which serves as a proof of concept for the proposed method. An approach similar in spirit to our work was presented for single-view video sequences coded with frames in MR [18]. In this proposal, we use available depth information to register the correspondences among views.

II. Resolution Recovery

Super-resolution (SR) techniques combine low-resolution images to obtain a high-resolution image [19]. Generally speaking, there are two main categories of algorithms: multi-image and example-based SR [20]. The first category consists of using several low-resolution images with subpixel misalignments to recover a high-resolution image, and the second category uses databases of low-resolution and high-resolution image pairs to correlate them with the low-resolution image to be super-resolved. Our method resembles the example-based SR methods such as that by Freeman *et al.* [20]. However, it does not use image databases. Instead, high-frequency details for low-resolution views are obtained from full-resolution views, which present high correlation with the target view image. That is, within the multiview setup with MR, images from high-resolution views offer the high frequency details to a particular low-resolution view image. The low-resolution images, furnished in these MR setups, are generally formed with common anti-aliasing subsampling, serving for both coding and complexity reduction.

Our goal is to make an estimate, $\hat{V}_n$, of the nth view's original full-resolution version, V_n. We define the low-resolution image of this view as V_n^D, which can be obtained by a simple process of downsampling V_n ($\downarrow M$, where M is the downsampling factor). This process consists of low-pass filtering, for anti-aliasing, and decimation, to reduce the image dimensions. If we upsample ($\uparrow M$) V_n^D, we obtain a low-pass version of V_n, defined here as V_n^L. The high-pass version of V_n, V_n^H, is the difference between V_n and V_n^L as follows:

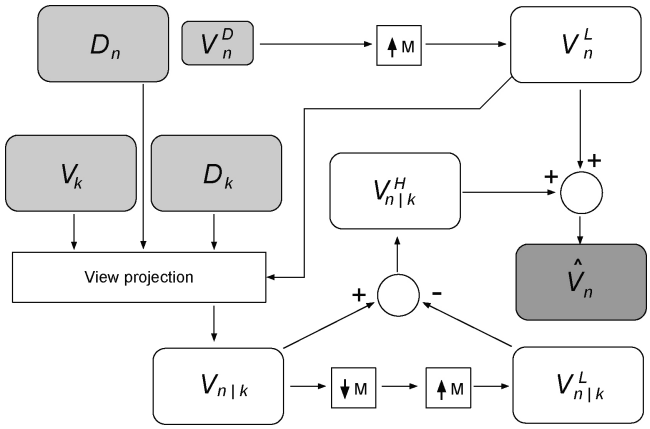

Fig. 1. SR approach for multiview images. A super-resolved image $\hat{V}_n$ is created from its low-resolution version, V_n^D, a neighboring full-resolution view, V_k, and the depth information for each of these views, D_n and D_k.

$$V_n^H = V_n - V_n^L. \tag{1}$$

Hence, by estimating V_n^H from one or more adjacent views, we may super-resolve V_n^D and determine $\hat{V}_n$. We will explain the method for one view, and then extrapolate for more views. Considering that we have another available view in full resolution, V_k, our method follows three basic steps. The first one is to directly project V_n from V_k, defined here as $V_{n|k}$. In this paper, we assume we are dealing with a multiview video-plus-depth (MVD) framework, such that the depth information from each view, D_n and D_k, is available. This information is very important for the view projection, as it indicates the pixel correspondences between views.

$V_{n|k}$, in itself, could be our estimate $\hat{V}_n$, but we would be neglecting important available information from the target view, V_n^L. So, in the next step, we obtain the high frequency information of $V_{n|k}$, defined as $V_{n|k}^H$. In order to do that, we downsample and upsample $V_{n|k}$, which generates its low-pass version $V_{n|k}^L$, and then take the difference between the two, in the same manner as (1)

$$V_{n|k}^H = V_{n|k} - V_{n|k}^L. \tag{2}$$

In the final step, we add $V_{n|k}^H$ to V_n^L to obtain the final estimate $\hat{V}_n$ as follows:

$$\hat{V}_n = V_n^L + V_{n|k}^H. \tag{3}$$

Fig. 1 illustrates our general method for recovering high frequency information for low-resolution views, which will be detailed next.

A. View Projection

The first step in our resolution-recovery method is to project view k onto view n. In other words, the image points in V_n

need to be projected from their corresponding positions in view k, in order to generate $V_{n|k}$. Therefore, point correspondence between views is needed.

Our resolution-recovery method assumes that the depth information of each view is available, in what is known as the MVD format [16]. In this manner, not only are the video sequences captured by cameras from different view points and synchronized in time but also each view has its own depth map, which is a grayscale image containing normalized depth information with respect to a global coordinate system. Having the depth maps and the camera calibration parameters, image points can be projected onto the 3-D space and back, yielding pixel correspondences among multiple views. The MVD format is presented as an alternative to multiview video systems, as new views can be interpolated from the available views and their depths [3]. Furthermore, depth maps can be efficiently encoded, representing low overhead [21].

In order to establish correspondences between views n and k at full resolution, we assume a pinhole camera model and project point coordinates from view n onto the 3-D space. Then, 3-D points are reprojected onto the view k, establishing a correspondence. The intrinsic parameters $\mathbf{A}$ of size 3×3, rotation matrix $\mathbf{R}$ of size 3×3, translation vector $\mathbf{t}$ of size 3×1, and corresponding depth map D of camera n are used to project pixel location (u, v) into world coordinates (x, y, z) [21] as follows:

$$[x, y, z]^T = \mathbf{R_n}\mathbf{A_n}^{-1}[u, v, 1]^{TD_n}(u, v) + \mathbf{t_n}. \tag{4}$$

The 3-D points are then reprojected onto view k, yielding coordinates (u', v') as follows:

$$[u' * w', v' * w', w']^T = \mathbf{A_k}\mathbf{R_k}^{-1}\left\{[x, y, z]^T - \mathbf{t_k}\right\}. \tag{5}$$

After establishing correspondences between views n and k, there are still two important aspects to be considered. First, the depth maps are susceptible to errors, derived from occlusions, depth imprecision, or compression quantization, which may generate incorrect correspondences. Second, the depth maps may generate correspondences with subpixel precision. In this case, pixel interpolation becomes necessary.

The proposed solution was to establish a consistency check using the available pair of depth maps. Based on (4) and (5), as depth map D_n points position (u, v) to position (u', v') in depth map D_k, we find the closest integer positions to position (u', v') as follows:

$$\begin{aligned} p_1 &= (\lfloor u' \rfloor, \lfloor v' \rfloor) \\ p_2 &= (\lceil u' \rceil, \lfloor v' \rfloor) \\ p_3 &= (\lfloor u' \rfloor, \lceil v' \rceil) \\ p_4 &= (\lceil u' \rceil, \lceil v' \rceil) \end{aligned} \tag{6}$$

where $\lfloor . \rfloor$ is the *floor* operation and $\lceil . \rceil$ is the *ceil* operation. Between points $\{p_1, p_2, p_3, p_4\}$, we select the one closest to (u', v'), and follow its corresponding position (u'', v'') back in D_n, based on D_k. If (u'', v'') falls inside a one-pixel radius, $\{p_1, p_2, p_3, p_4\}$ are used to calculate $V_{n|k}(u, v)$ via bilinear interpolation. If (u'', v'') falls outside of this radius, then no interpolation is done and the value of $V_n^L(u, v)$ is used for $V_{n|k}(u, v)$. In this manner, depth inconsistencies are detected, and erroneous high-frequency information from adjacent views are not added to the projection. Fig. 2 illustrates the consistency check between depth maps.

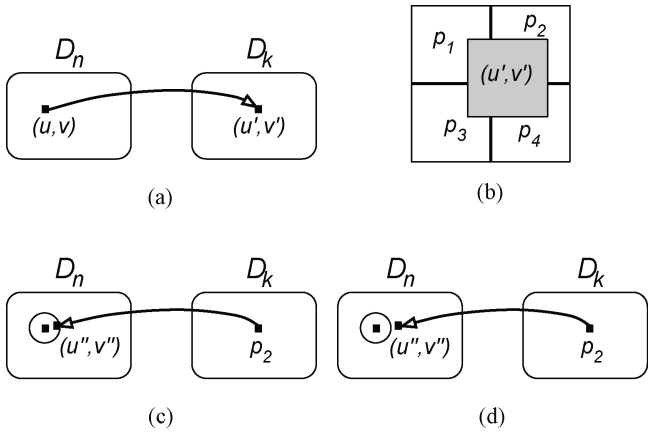

Fig. 2. Consistency check for depth maps. (a) Projection from position (u, v) to (u', v'). (b) Floating point position (u', v') in D_k and closest integer positions p_1, p_2, p_3, and p_4. (c) Reprojection of p_2 from D_k to D_n, falling inside a one-pixel radius around (u, v), and thus yielding a valid projection. (d) Reprojection of p_2 from D_k to D_n, falling outside the radius and yielding an invalid projection.

B. Multiple Views

If there is more than one full-resolution view available, there are more high-frequency candidates from which to select, strengthening the estimate of V_n. In this paper, we chose to sum the candidates from all available full-resolution views, using weights to ponder them. First, we generate $V_{n|i}^H$, for $i = \{1, \ldots, N\}$ being the set of indices of the full-resolution reference views. Next, for all positions (u, v) in $\hat{V}_n$, we calculate $d_{n,i}$, which is the Euclidean distance between points (u, v) and (u'', v''), as defined in Section II-A. Furthermore, we calculate $m_{n,i}$, which is a mask that reflects the consistency check described in Section II-A, as follows:

$$m_{n,i} = \begin{cases} 1, & d_{n,i} < 1 \\ 0, & d_{n,i} \geqslant 1. \end{cases} \tag{7}$$

The final estimate of V_n for each position (u, v), for N views, is given by

$$\hat{V}_n(u, v) = V_n^L(u, v) + \left(\sum_{i=1}^{N} \frac{V_{n|i}^H(u, v) m_{n,i}}{d_{n,i}}\right) / \left(\sum_{i=1}^{N} \frac{m_{n,i}}{d_{n,i}}\right). \tag{8}$$

In this manner, for each view i, we use the degree of consistency between D_n and D_i (represented by $d_{n,i}$) to ponder how much high frequency from $V_{n|i}^H$ should be used in estimating V_n. The lower the value of $d_{n,i}$, the higher the consistency and, therefore, the higher the weight given to $V_{n|i}^H$. If the point does not pass the consistency check, $m_{n,i} = 0$, and the point is not considered in the final estimate.

III. Experimental Results

The proposed method was tested on a series of multiview images and sequences, both for synthetic and real image data, under several compression conditions. First, we tested our SR

method with the original color images and their associated depths, without any compression. Next, the method was tested with unmatched downsampling and upsampling filters, in order to evaluate the effect for the overall results. Third, the method was tested for images and depth maps encoded with MR, for a wide range of bit-rates. Last, in order to assess a more complete bit-rate distribution between color images and depths for the various content, the depth images were encoded at a series of quantization parameters (QPs) different from those of the color images. In order to evaluate the coding results, we present rate-distortion (RD) curves and tables and used the popular Bjøntegaard metric [22] for calculating average gains between curves.

Table I shows the peak signal-to-noise ratio (PSNR) for the luma component of two versions of the low-resolution image, an interpolated (or upsampled) version and a super-resolved version using the proposed method. A 6-tap Lanczos interpolation filter was used for downsampling and upsampling, for factors $M = 2$ and $M = 4$. We used the following synthetic stereo images for our tests: *Barn1*, *Barn2*, *Bull*, *Cones*, *Map*, *Poster*, *Sawtooth*, *Teddy*, and *Venus* [1], where the right view was adopted as low resolution and the left view was maintained as full resolution. As for the real image data, we used frame 0 of multiview sequences *Ballet*, *Breakdancers* [23], *Pantomime*, *Dog* [24], and *Poznan Street* [25]. The depth maps for the *Pantomime* and *Dog* images were computed using a graph-cuts stereo-correspondence software [26]. Due to lack of high-frequency content in the original images, *Ballet*, *Breakdancers*, *Pantomime*, *Dog*, and *Poznan Street* were resized to 512×384, 256×192, 640×480, 640×480, and 960×544, respectively, for these evaluations. For *Ballet* and *Breakdancers*, view 1 was chosen as the low-resolution view and was super-resolved with the aid of a full-resolution view 2. For *Pantomime* and *Dog*, view 39 was chosen as the low-resolution view and was super-resolved with the aid of a full-resolution view 40. For *Poznan Street*, view 4 was chosen as the low-resolution view and was super-resolved with the aid of a full-resolution view 5.

Table I reveals quality improvement for all test images, both for factors $M = 2$ and $M = 4$. In the worst case (sequence *Dog*, $M = 2$), there is a 1.06 dB improvement, and in the best case (sequence *Barn2*, $M = 4$), there is an 9.28 dB improvement.

In Table II, results are presented for the low-resolution view from the multiview sequences *Ballet*, *Breakdancers*, *Dog*, and *Pantomime*. *Ballet* and *Breakdancers* were super-resolved with the aid of multiple full-resolution views 0 and 2, *Pantomime* and *Dog*, with views 38 and 40, and *Poznan Street*, with views 3 and 5, as described in Section II-B. As in Table I, frame 0 was tested for all sequences. The usage of multiple views for SR achieves gains between the range of 0.9 dB and 2.81 dB (*Breakdancers*, $M = 2$, and *Pantomime*, $M = 4$, respectively), over the single-view SR results described in Table I. Such gains hail from the availability of multiple high-frequency candidates that are either pondered or used to complement an inconsistent point projection, thus strengthening the SR estimates. Fig. 3 presents details of the interpolated, super-resolved, and original versions of the *Pantomime* sequence, where a great quality improvement can be seen on the clown's face, hat, and clothes.

TABLE I
LUMA PSNR RESULTS FOR UNCOMPRESSED SEQUENCES COMPARING SR AND INTERPOLATION

Sequence	M	Interpolated (dB)	Proposed SR (dB)
Barn1	2	27.76	35.83
	4	24.93	33.99
Barn2	2	31.06	38.40
	4	27.42	36.70
Bull	2	32.46	37.96
	4	28.49	35.79
Map	2	28.00	31.20
	4	20.75	27.92
Poster	2	26.46	33.93
	4	22.78	31.91
Sawtooth	2	28.32	33.72
	4	24.57	31.89
Venus	2	28.63	35.61
	4	25.15	33.37
Cones	2	28.88	33.04
	4	24.93	30.32
Teddy	2	30.38	34.20
	4	26.26	31.41
Ballet	2	34.03	36.34
	4	29.84	33.61
Breakdancers	2	35.53	39.09
	4	30.12	36.62
Pantomime	2	36.02	38.59
	4	28.11	34.40
Dog	2	35.48	36.54
	4	27.04	32.37
Poznan Street	2	32.10	35.78
	4	26.57	32.44

(a) 28.11 dB (b) 37.21 dB (c)

Fig. 3. Details of the results for the *Pantomime* view 39, frame 0. (a) Downsampled and upsampled by 4 with a Lanczos filter. (b) Downsampled by 4 and super-resolved with a Lanczos filter, using two full-resolution views. (c) Original. Results are best seen on a screen.

In reference to previous work [17], the relative gains between the proposed SR and the interpolated versions (both for single and multiple view) are significantly better. In the worst case, the previous work reports a gain of 0.6 dB (*Ballet*, single-view reference, $M = 2$), while our method offers a 2.01 dB gain for the same sequence. In the best case, the previous work reports a gain of 2.9 dB (*Breakdancers*, multiple-view references, $M = 4$), while our method offers a 7.73 dB gain for the same sequence. These gains can be attributed to two fundamental differences between the proposed and previous methods: an improved consistency check for the depth maps and a superior high-frequency extraction methodology, where the adjacent view is projected prior to frequency decomposition.

TABLE II
LUMA PSNR RESULTS FOR UNCOMPRESSED SEQUENCES COMPARING SR WITH SINGLE VIEW AND MULTIPLE VIEWS

Sequence	M	Single View (dB)	Multiple Views (dB)
Ballet	2	36.34	38.02
	4	33.61	35.53
Breakdancers	2	39.09	39.99
	4	36.62	37.85
Pantomime	2	38.59	40.91
	4	34.40	37.21
Dog	2	36.54	38.12
	4	32.37	33.38
Poznan Street	2	35.78	37.14
	4	32.44	34.21

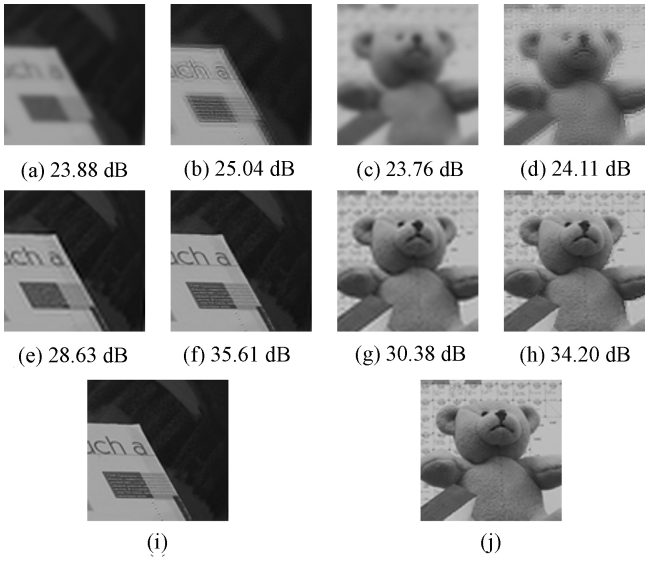

Fig. 4. Details of the results for the *Venus* and *Teddy* images. (a) *Venus*, downsampled with a bicubic filter and upsampled with a Lanczos filter. (b) *Venus*, downsampled with a bicubic filter and super-resolved with a Lanczos filter. (c) *Teddy*, downsampled with a bicubic filter and upsampled with a Lanczos filter. (d) *Teddy*, downsampled with a bicubic filter and super-resolved with a Lanczos filter. (e) *Venus*, downsampled and upsampled with a Lanczos filter. (f) *Venus*, downsampled and super-resolved with a Lanczos filter. (g) *Teddy*, downsampled and upsampled with a Lanczos filter. (h) *Teddy*, downsampled and super-resolved with a Lanczos filter. (i) *Venus*, original. (j) *Teddy*, original. Results are best seen on a screen.

It is possible that our SR method is applied to a multiview sequence in MR where the downsampling method is unknown. In order to evaluate the effect of using different (unmatched) downsampling and upsampling filters, we conducted two experiments. In the first one, the low-resolution image is created using a bicubic downsampling filter, while interpolation and SR employ the previously mentioned 6-tap Lanczos filtering. In the second experiment, Lanczos filtering is used for downsampling, interpolation, and SR (equivalent to Table I). A downsampling and upsampling factor $M = 2$ was applied in all tests. The same view and frame numbers used in Table I were applied.

From the results in Table III, it can be seen that using unmatched downsampling and upsampling filters may degrade the performance of the proposed method. For instance, sequence *Sawtooth* presents a gain of 5.4 dB between the super-resolved and interpolated versions in the matched-filter case, and a 0.88 dB gain in the unmatched-filter case. This is due to the fact that the proposed method assumes matched downsampling and upsampling filters for the high-frequency extraction of the projected view, as depicted in Fig. 1. Nonetheless, the proposed SR offers gains for most tested sequences, even when applying unmatched downsampling and upsampling filters. The only quality loss is for sequence *Dog*, where there is a 0.3 dB drop.

TABLE III
LUMA PSNR RESULTS FOR UNMATCHED AND MATCHED DOWNSAMPLING AND UPSAMPLING FILTERS

Sequence	Method	Bicubic/Lanczos (dB)	Lanczos/Lanczos (dB)
Barn1	Interp.	23.98	27.76
	SR	25.73	35.83
Barn2	Interp.	25.47	31.06
	SR	26.37	38.40
Bull	Interp.	26.20	32.46
	SR	26.76	37.96
Map	Interp.	19.50	28.00
	SR	19.52	31.20
Poster	Interp.	21.94	26.46
	SR	23.22	33.93
Sawtooth	Interp.	23.37	28.32
	SR	24.25	33.72
Venus	Interp.	23.88	28.63
	SR	25.04	35.61
Cones	Interp.	23.42	28.88
	SR	24.01	33.04
Teddy	Interp.	23.76	30.38
	SR	24.11	34.20
Ballet	Interp.	27.72	34.03
	SR	27.87	36.34
Breakdancers	Interp.	27.39	35.53
	SR	27.44	39.09
Pantomime	Interp.	26.48	36.02
	SR	26.48	38.59
Dog	Interp.	24.66	35.48
	SR	24.36	36.54
Poznan Street	Interp.	25.11	32.10
	SR	25.29	35.78

Fig. 4 presents details of the interpolated, super-resolved, and original versions of sequences *Venus* and *Teddy*, applying matched and unmatched downsampling and upsampling filters. Fig. 4(a)–(d) depicts the unmatched-filter case and Fig. 4(e)–(h) depicts the matched-filter case. It can be seen that the high-frequency extraction does not work for the unmatched-filter case as efficiently as in the matched-filter case, but there is still a quality improvement. As for the matched-filter case, significant quality improvements are noticeable.

Next, all images were encoded with the H.264/AVC reference software JM 17.2 [27] in Intra mode, with MR. Intra coding, in this case, serves as a proof of concept, demonstrating performance under a wide range of bit-rates. The proposed method does not make assumptions regarding the used prediction mode and may thus be extended to Inter coding as well. For synthetic sequences, the left view was encoded at

TABLE IV
AVERAGE LUMA PSNR GAIN BETWEEN SR AND INTERPOLATION

Sequence	M	Average PSNR Gain (dB)
Barn1	2	3.26
	4	4.10
Barn2	2	2.43
	4	3.76
Bull	2	1.87
	4	3.24
Map	2	1.30
	4	4.79
Poster	2	3.76
	4	5.22
Sawtooth	2	2.46
	4	3.82
Venus	2	3.44
	4	4.56
Cones	2	1.42
	4	2.17
Teddy	2	1.35
	4	2.00
Ballet	2	1.19
	4	2.18
Breakdancers	2	0.81
	4	2.42
Pantomime	2	1.39
	4	4.78
Dog	2	−0.16
	4	3.76
Poznan Street	2	1.50
	4	3.49

full resolution and the right view was downsampled by factors $M = 2$ and $M = 4$ prior to encoding. Frame 0 was tested for all real image data. For sequences *Ballet* and *Breakdancers*, view 2 was encoded at full resolution while view 1 was downsampled by factors $M = 2$ and $M = 4$. For sequences *Dog* and *Pantomime*, view 40 was encoded at full resolution while view 39 was downsampled by factors $M = 2$ and $M = 4$. For sequence *Poznan Street*, view 5 was encoded at full resolution while view 4 was downsampled by factors $M = 2$ and $M = 4$. Rate-control and RD optimization were disabled and the QPs in the set QP = {7, 12, 17, 22, 27, 32, 37, 42, 47} were used for both color and depth. In order to calculate average gains between RD curves, four quantization points were considered QP = {22, 27, 32, 37}.

Table IV presents the average PSNR gain for these sequences, relative to only interpolating the low-resolution view. In this table, both the color and depth images were encoded using the same QPs. The rate was considered as the total bit-rate of the system (stereo color and depth images). This decision is based on an assumption of a free-viewpoint television context, wherein the user might choose to see the low-resolution view, which is the one being improved with aid of an adjacent full-resolution view. Since our SR method does not require extra information from the encoder, the bit-rates for the interpolated image and the super-resolved one are the same. Figs. 5 and 6 present the RD performance for images *Venus* and *Breakdancers*.

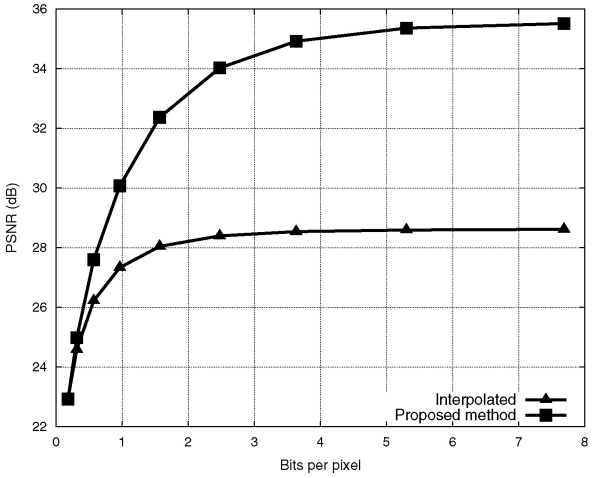

Fig. 5. RD performance for the *Venus* right image, compressing both color and depth images.

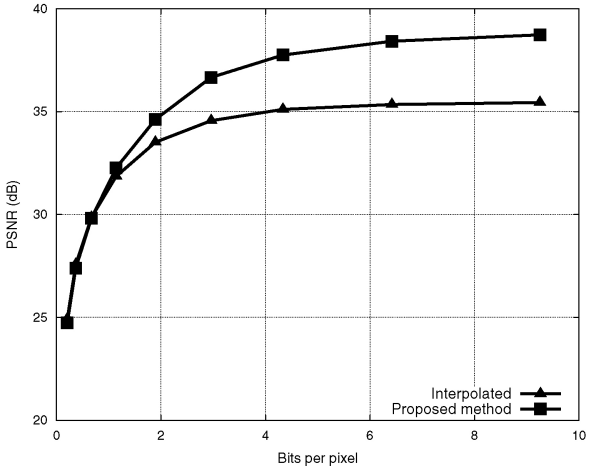

Fig. 6. RD performance for the *Breakdancers* view 1, frame 0, compressing both color and depth images.

Table IV shows that, except for sequence *Dog* with $M = 2$, our method presents overall PSNR gain for the test images. However, this gain is not homogeneous, as Figs. 5 and 6 demonstrate. At low bit-rates, there is little to improve, as both the low-resolution right image, the full-resolution left image, and the associated depth images are highly distorted after coding. As the bit-rate increases, the performance of the SR method significantly improves, presenting PSNR differences as high as 7 dB (Fig. 5), until it appears to saturate. That is, above a certain bit-rate value, there is no extra rate that can improve the quality of both the interpolated and the super-resolved images. This behavior is due to the fact that if an image is encoded at a low resolution, it already has an upper quality bound that cannot be surpassed. In the case of interpolation, this limit is the distortion from downsampling and upsampling, and in the case of the SR method, it is the result of super-resolving this downsampled and upsampled image without any coding considerations.

Next, we assessed the effects of choosing different bit-rate distributions between color images and depth by selecting QP for the depth images different from those of the color images.

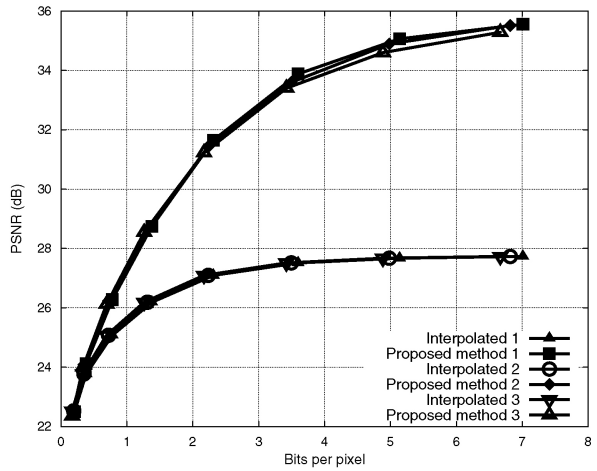

Fig. 7. RD performance for the *Barn1* right image, for depth images coded with QPs different from those of the color images.

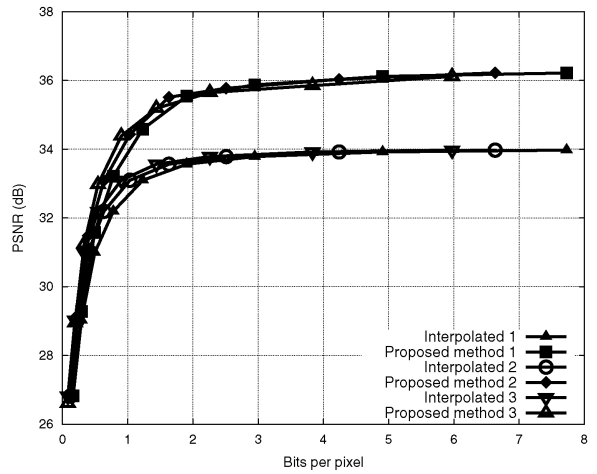

Fig. 8. RD performance for the *Ballet* view 1, frame 0, for depth images coded with QPs different from those of the color images.

Let us define cQP as the QP used for the color images and dQP as the QP used for the depth images. We compared the effect of choosing

$$\begin{aligned} 1: &\quad d\text{QP} = c\text{QP} - 6 \\ 2: &\quad d\text{QP} = c\text{QP} \\ 3: &\quad d\text{QP} = c\text{QP} + 6. \end{aligned} \tag{9}$$

The second case corresponds to the tests presented in Table IV.

Figs. 7–10 present the results for synthetic and real images *Barn1*, *Ballet*, *Bull*, and *Pantomime*, respectively, with testing conditions of (9). These images represent two patterns: Figs. 7 and 8 show cases where the RD performance changes very little according to the dQP variation, and Figs. 9 and 10 show cases where choosing $d\text{QP} < c\text{QP}$ renders worse RD performance, both for interpolating the low-resolution view and for super-resolving it. In terms of using interpolation and using our method, Figs. 7–10 show that it is slightly better to choose dQP = cQP+6. Nevertheless, it is always advantageous to use SR over simple interpolation.

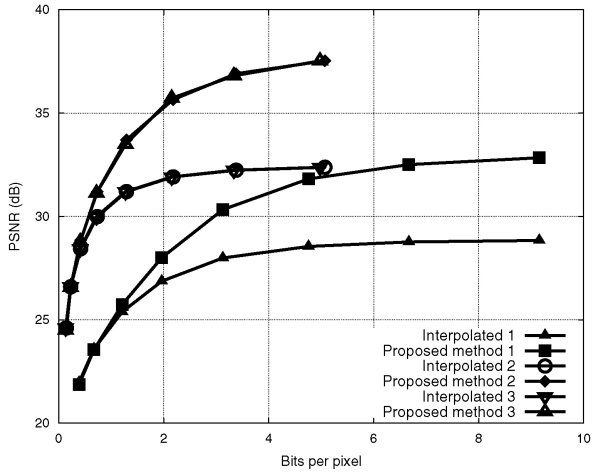

Fig. 9. RD performance for the *Bull* right image, for depth images coded with QPs different from those of the color images.

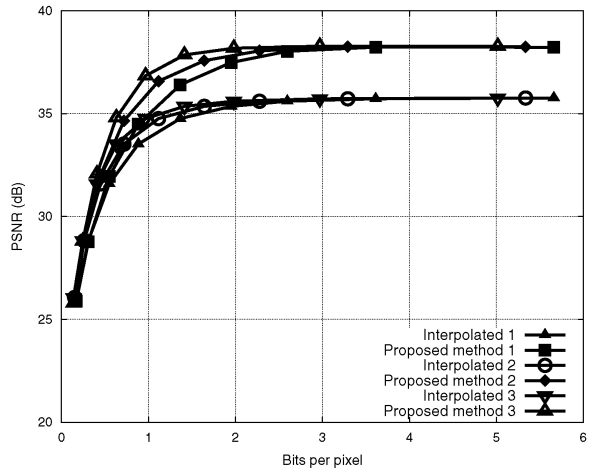

Fig. 10. RD performance for the *Pantomime* view 39, frame 0, for depth images coded with QPs different from those of the color images.

IV. Conclusion

In this paper, we presented a novel SR technique to improve low-resolution views in MR multiview-plus-depth systems. Low-resolution views were enhanced with the high-resolution information from neighboring full-resolution views. The correspondence between views was carried out with the depth information from all views, and occlusions and depth mismatches were corrected based on depth correspondences between views. Tests with synthetic and real data images showed that our method renders great improvement over interpolation of the low-resolution views, outperforming previous work [17]. Tests were conducted for the same sequences coded with the H.264/AVC codec, still presenting gains over interpolation. For both of these methods, there was a quality saturation, where the increase in bit-rate offered little quality improvement. The effect of applying unmatched downsampling and upsampling filters were also evaluated, showing that gains were achievable even with unmatched filter pairs. Furthermore, it was shown that the quantization parameter for the depth images can be increased in relation to the quantization parameter for the color images, offering a slight improvement in the overall

RD performance. Future works involve assessing the effect of using prediction modes other than Intra coding over the proposed method.

REFERENCES

[1] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *Int. J. Comput. Vision*, vol. 47, nos. 1–3, pp. 7–42, Apr.–Jun. 2002.

[2] ITU-T and ISO/IEC JTC 1, "Advanced video coding for generic audiovisual service," document ITU-T Rec. H.264 and ISO/IEC 14496-10 (MPEG-4 AVC), 2010.

[3] A. Vetro, S. Yea, and A. Smolic, "Towards a 3D video format for auto-stereoscopic displays," *Proc. SPIE Conf. Applicat. Digit. Image Process.*, vol. 7073, pp. 70730F-1–70730F-10, Sep. 2008.

[4] B. Julesz, *Foundations of Cyclopean Perception*. Chicago, IL: Univ. Chicago Press, 1971.

[5] W. Tam, "Image and depth quality of asymmetrically coded stereoscopic video for 3D-TV," document JVT-W094, San Jose, CA, Apr. 2007.

[6] P. Aflaki, M. Hannuksela, J. Häkkinen, P. Lindroos, and M. Gabbouj, "Subjective study on compressed asymmetric stereoscopic video," in *Proc. Int. Conf. Image Process.*, Sep. 2010, pp. 4021–4024.

[7] G. Saygili, C. Gurler, and A. Tekalp, "Quality assessment of assymmetric stereo video coding," in *Proc. Int. Conf. Image Process.*, Sep. 2010, pp. 4009–4012.

[8] C. Fehn, P. Kauff, S. Cho, H. Kwon, N. Hur, and J. Kim, "Asymmetric coding of stereoscopic video for transmission over T-DMB," in *Proc. 3DTV-CON*, May 2007, pp. 1–4.

[9] H. Brust, G. Tech, K. Mueller, and T. Wiegand, "Mixed resolution coding with inter view prediction for mobile 3DTV," in *Proc. 3DTV-CON*, Jun. 2010, pp. 1–4.

[10] E. Ekmekcioglu, S. Worrall, and A. Kondoz, "Utilisation of downsampling for arbitrary views in multi-view video coding," *IEEE Signal Process. Lett.*, vol. 44, no. 5, pp. 339–340, Feb. 2008.

[11] A. Aksay, C. Bilen, E. Kurutepe, T. Ozcelebi, G. Akar, R. Civanlar, and A. Tekalp, "Temporal and spatial scaling for stereoscopic video compression," in *Proc. 14th Eur. Signal Process. Conf.*, Sep. 2006, pp. 4–8.

[12] Y. Chen, Y.-K. Wang, M. Gabbouj, and M. Hannuksela, "Regionally adaptive filtering for asymmetric stereoscopic video coding," in *Proc. IEEE Int. Symp. Circuits Syst.*, May 2009, pp. 2585–2588.

[13] H. Sawhney, Y. Guo, K. Hanna, and R. Kumar, "Hybrid stereo camera: An IBR approach for synthesis of very high resolution stereoscopic image sequences," in *Proc. SIGGRAPH*, 2001, pp. 451–460.

[14] D. Meegan, L. Stelmach, and W. Tam, "Unequal weighting of monocular inputs in binocular combination: Implications for the compression of stereoscopic imagery," *J. Exp. Psychol.: Appl.*, vol. 7, no. 2, pp. 143–153, 2001.

[15] P. Seuntiens, L. Meesters, and W. Ijsselsteijn, "Perceived quality of compressed stereoscopic images: effects of symmetric and asymmetric JPEG coding and camera separation," *ACM Trans. Applied Perception*, vol. 3, no. 2, pp. 96–109, Apr. 2006.

[16] P. Merkle, A. Smolic, K. Muller, and T. Wiegand, "Multi-view video plus depth representation and coding," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2007, pp. 201–204.

[17] D. Garcia, C. Dórea, and R. de Queiroz, "Super-resolution for multiview images using depth information," in *Proc. Int. Conf. Image Process.*, Sep. 2010, pp. 1793–1796.

[18] F. Brandi, R. de Queiroz, and D. Mukherjee, "Super-resolution of video using key frames and motion estimation," in *Proc. Int. Conf. Image Process.*, Oct. 2008, pp. 321–324.

[19] S. Baker and T. Kanade, "Limits on super-resolution and how to break them," *IEEE Trans. Patt. Anal. Mach. Intell.*, vol. 24, no. 9, pp. 1167–1183, Sep. 2002.

[20] W. Freeman, T. Jones, and E. Pasztor, "Example-based super-resolution," *IEEE Comput. Graph. Applicat.*, vol. 22, no. 2, pp. 56–65, Mar. 2002.

[21] P. Kauff, N. Atzpadin, C. Fehn, M. Mller, O. Schreer, A. Smolic, and R. Tanger, "Depth map creation and image based rendering for advanced 3DTV services providing interoperability and scalability," *Signal Process.: Image Commun.*, vol. 22, no. 2, pp. 217–234, Feb. 2007.

[22] G. Bjøntegaard, "Calculation of average PSNR differences between RD-curves," document VCEG-M33, Austin, TX, Apr. 2001.

[23] C. Zitnick, S. Kang, M. Uyttendaele, S. Winder, and R. Szeliski, "High-quality video view interpolation using a layered representation," in *Proc. SIGGRAPH*, 2004, pp. 600–608.

[24] Nagoya University. *FTV Test Sequences* [Online]. Available: http://www.tanimoto.nuee.nagoya-u.ac.jp

[25] M. Domański, T. Grajek, K. Klimaszewski, M. Kurc, O. Stankiewicz, J. Stankowski, and K. Wegner, "Poznań multiview video test sequences and camera parameters," document MPEG 2009/M17050, ISO/IEC JTC1/SC29/WG11, Xian, China, Oct. 2009.

[26] V. Kolmogorov and R. Zabih, "Computing visual correspondence with occlusions using graph cuts," in *Proc. IEEE Int. Conf. Comp. Vision*, vol. 2. Jul. 2001, pp. 508–515.

[27] *JM H.264/AVC Reference Software* [Online]. Available: http://iphome.hhi.de/suehring/tml

Diogo C. Garcia (S'10) received the Electrical Engineer and M.S. degrees in electrical engineering from the University of Brasilia, Brasilia, Brazil, in 2006 and 2008, respectively, where he is currently pursuing the Ph.D. degree.

He is currently an Adjunct Professor with the Gama Faculty of the University of Brasilia. His current research interests include image and video coding, super resolution, and multiview and 3-D processing.

Camilo Dorea received the B.S. degree from the University of Brasilia, Brasilia, Brazil, in 1997, the M.S. degree from the University of Maryland, College Park, in 1999, both in electrical engineering, and the Ph.D. degree in telecommunications from the Technical University of Catalonia, Barcelona, Spain, in 2007.

From 2007 to 2008, he was with Thomson Corporate Research, Princeton, NJ. In 2009, he was with the Department of Computer Science, University of Brasilia, where he is currently an Assistant Professor. His current research interests include video segmentation and analysis, video coding, and multiview and 3-D processing.

Ricardo L. de Queiroz (SM'99) received the Electrical Engineer degree from the University of Brasilia, Brasilia, Brazil, in 1987, the M.S. degree from the University of Campinas, Campinas, Brazil, in 1990, and the Ph.D. degree from the University of Texas at Arlington, Arlington, in 1994, all in electrical engineering.

From 1990 to 1991, he was a Research Associate with the DSP Research Group, University of Brasilia. He joined Xerox Corporation, Webster, NY, in 1994, where he was a Research Staff Member until 2002. From 2000 to 2001, he was an Adjunct Faculty Member with the Rochester Institute of Technology, Rochester, NY. In 2003, he was with the Department of Electrical Engineering, University of Brasilia. In 2010, he became a Full Professor with the Department of Computer Science, University of Brasilia. He has published over 150 articles in journals and conferences and contributed chapters to books as well. He also holds 46 issued patents. His current research interests include image and video compression, multirate signal processing, and color imaging.

Dr. de Queiroz is an elected member of the IEEE Signal Processing Society's Multimedia Signal Processing (MMSP) Technical Committee and a former member of the Image, Video, and Multidimensional Signal Processing Technical Committee. He was a past editor of the *EURASIP Journal on Image and Video Processing*, the IEEE SIGNAL PROCESSING LETTERS, the IEEE TRANSACTIONS ON IMAGE PROCESSING, and the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY. He is appointed an IEEE Signal Processing Society Distinguished Lecturer for the 2011–2012 term. He was actively involved with the Rochester Chapter of the IEEE Signal Processing Society, where he was the Chair and organized the Western New York Image Processing Workshop from its inception until 2001. He is now helping organizing IEEE SPS Chapters in Brazil and just founded the Brasilia IEEE SPS Chapter. He was the General Chair of MMSP'2009 and ISCAS'2011, and is the General Chair of SBrT'2012. He was also part of the organizing committee of ICIP'2002. He is a member of the Brazilian Telecommunications Society and of the Brazilian Society of Television Engineers.