SUPER-RESOLUTION FOR MULTIVIEW IMAGES USING DEPTH INFORMATION

Diogo C. Garcia[1], *Camilo Dórea*[2], *Ricardo L. de Queiroz*[2]

[1]Departamento de Engenharia Elétrica, [2]Departamento de Ciência da Computação
Universidade de Brasília, DF, Brasil
{diogo, camilo, queiroz}@image.unb.br

ABSTRACT

The joint usage of low– and full-resolution images in multiview systems provides an attractive opportunity for data size reduction while maintaining good quality in 3D applications. In this paper we present a novel application of a super-resolution method for usage within a mixed resolution multiview setup. The technique borrows high-frequency content from neighboring full resolution images to enhance particular low-resolution views. Occlusions are handled through matching of low-resolution images. Both the stereo and the more general multiview cases are considered using the multiview video-plus-depth format. Results demonstrate significant gains in PSNR and in visual quality for test sequences.

Index Terms— Super-resolution, Mixed Resolution, Multiview Video.

1. INTRODUCTION

Current advances in multiview video acquisition and display technology have made 3D video one of the most promising applications in video entertainment. They offer new user experiences such as auto-stereoscopic displays (3D TV) and free-viewpoint video [1], which may branch themselves into a whole series of applications, such as mobile 3D TV, immersive teleconferencing, among others. Along with these diverse opportunities in video entertainment innovation, comes the burden of massive growth in data size and processing complexity. For instance, a simple camera stereo pair doubles the amount of raw data to be compressed, while further demands for view synthesis at the decoder side may increase the video processing tremendously. Nevertheless, subjective aspects of the human visual system suggest that good data size compression can be achieved by reducing the resolution of particular images in stereo pairs while maintaining a good subjective quality in the 3D experience.

Previous psychological and physiological studies [2] observed that the human visual system developed clever ways to compensate for some forms of asymmetric qualities in stereo vision, in what is known as suppression theory. It has been shown that when one of the views is low-pass filtered, the binocular image's quality is not affected. The scene's apparent depth remains the same and its sharpness is maintained with high-frequency information from the other view. On the other hand, if one of the views has a higher quantization level during compression, blocking artifacts can reduce the binocular image's quality [3].

Suppression theory's findings suggest that a mixed resolution approach to stereo and multiview compression can greatly reduce the amount of data to be compressed, without incurring in subjective quality reduction. Several mixed resolution stereo coding frameworks for small devices such as mobile phones have been proposed [4] [5], coding one of the views entirely at a lower resolution. Objective gains in asymmetric coding for different down-sampling ratios at low bit rates have also been reported [6]. The addition of temporal scalability to a mixed resolution framework has been investigated [7], as well as the prediction of macroblocks of the low-resolution view directly from the high-resolution view [8], which eliminates the computational burden of sub-sampling reference frames. The generation of a high-resolution synthesis of the low-resolution sequence at the decoder was developed, using image-based rendering techniques [9].

The majority of the previous works do not compensate for the quality differences between views, originating two fundamental shortcomings. First, the techniques may not be suitable when the viewer's dominant eye is not the same as the system's high-resolution view [3]. Second, they may not be appropriate for the monoscopic nature of free-viewpoint television. For instance, quality reduction will be perceptible if the user happens to select the low-resolution view.

In order to circumvent these problems, we propose an approach to increase the objective quality of the low-resolution views, using the high-frequency information from neighboring views. The technique is based on a previous super-resolution work [10], which was applied to single-view video sequences coded with mixed resolution. The idea is to upsample the low-resolution sequence to its original dimensions and obtain the high-frequency information from neighboring views. In our work, we take advantage of the fact that depth information may be available or computed, taking care of

Work supported by HP Brasil.

the registration process of finding correspondences between views.

2. SUPER-RESOLUTION FOR MULTIVIEW

Super-resolution techniques are aimed at increasing an image's apparent resolution. For such, existing details in correlated images may be used to enhance, or super-resolve, an image. Within the multiview setup with mixed resolution, the images from adjacent views serve as the detail source used in super-resolving a particular low-resolution image.

An illustrative diagram of our super-resolution approach is shown in Fig. 1 for two views. The low-resolution image of the n-th view V_n is up-sampled to full resolution V_n^L and super-resolved with the aid of an image of an adjacent view V_{n+1}. Note that the adjacent image in this case is available at the original resolution and may be decomposed into low– and high-frequency images, V_{n+1}^L and V_{n+1}^H, respectively. These high-frequency components are used to recover the missing details V_n^H and to form the super-resolved image as $\hat{V}_n = V_n^L + V_n^H$.

The low-resolution image V_n is obtained by down-sampling ($\downarrow M$) an original image. This process includes low-pass filtering, which removes high-frequency content, and decimation, which reduces the image dimensions. Likewise, a low-frequency version of the adjacent image V_{n+1}^L may be formed by down-sampling and then up-sampling ($\uparrow M$) V_{n+1}. The difference between the original image and its low-frequency version is referred to as the high-frequency image $V_{n+1}^H = V_{n+1} - V_{n+1}^L$.

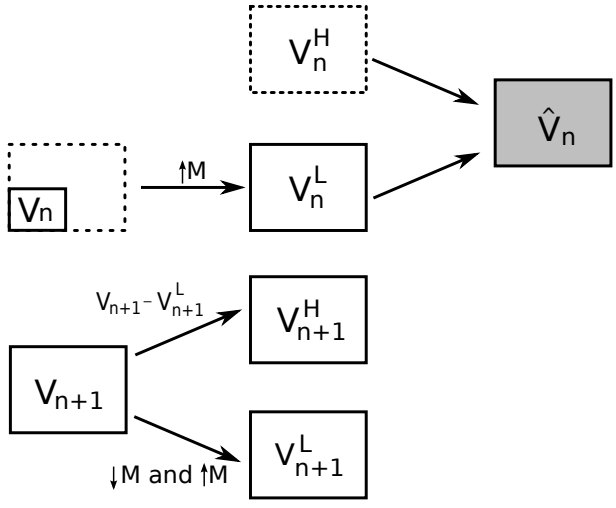

Fig. 1. Super-resolution approach for multiview images. An up-sampled V_n^L and high-frequency content V_{n+1}^H from the adjacent image are used to estimate V_n^H and form super-resolved image $\hat{V}_n$.

In the following sub-sections, we discuss the establishment of correspondences among various views and the super-resolution process for the stereo pair case and for the more general multiview case.

2.1. Calculation of view correspondence

In our approach, the multiview sequences are considered in the popular multiview video-plus-depth (MVD) format. The format consists of the multiple video sequences captured by synchronized cameras from different view points and corresponding depth maps. The depth maps are gray-scale images containing per pixel normalized depth information with respect to a global coordinate system. Along with the camera calibration parameters, depth maps can be used to project an image point into 3D space and are, thus, extremely useful in yielding correspondences between image points among multiple views. Furthermore, in terms of transmission requirements, depth maps can be efficiently coded, representing low overhead [11].

The super-resolution technique requires the establishment of correspondences between the low-frequency frames at full resolution: V_n^L and V_{n+1}^L. The well known pin-hole camera model is used to project point coordinates from a reference image onto their location in 3D space. Then, 3D points are re-projected onto the adjacent image, establishing a correspondence. The intrinsic parameters $\mathbf{A}$, rotation matrix $\mathbf{R}$, translation vector $\mathbf{t}$ and corresponding depth map D of camera n are used to project pixel location (u, v) into world coordinates (x, y, z) [11]:

$$(x, y, z)^T = \mathbf{R_n}\mathbf{A_n}^{-1}(u, v, 1)^T D_n(u, v) + \mathbf{t_n}. \quad (1)$$

Re-projection onto camera $n + 1$ yields target coordinates (u', v') from:

$$(u' * w', v' * w', w')^T = \mathbf{A_{n+1}}\mathbf{R_{n+1}}^{-1}[(x, y, z)^T - \mathbf{t_n}]. \quad (2)$$

2.2. Super-resolution for a stereo pair

Once correspondences have been established between pairs of views, high-frequency content from an adjacent view may be transferred to the low-resolution image. However, the previously described correspondence calculation is subject to occlusions between views. Furthermore, errors during depth map estimation, particularly around depth discontinuities, may also lead to invalid correspondences.

The super-resolution technique for stereo camera pairs accounts for these problems by requiring a sufficiently good match between correspondences among the low-frequency frames before updating with high-frequency components. Firstly, we define $V_{n+1}'^H$ as the bilinearly interpolated image intensities of V_{n+1}^H due to the use of non-integer re-projection results (u', v'), and SAD_{n+1}^L as the sum of absolute differences between square blocks of dimension w centered around locations (u, v) and (u', v') of the low-frequency image intensities V_n^L and $V_{n+1}'^L$, respectively. The super-resolved image

is thus given by:

$$\hat{V}_n(u,v) = V_n^L(u,v) + V_n^H(u,v) \quad (3)$$

and

$$V_n^H(u,v) = \begin{cases} V_{n+1}'^H(u',v') & \text{if } \mathrm{SAD}_{n+1}^L <= t \\ 0 & \text{else} \end{cases} \quad (4)$$

The use of SAD against a threshold t forms a low-frequency matching test. If matches are deemed inadequate, no high-frequency updating is attempted. Low-frequency matching based on SAD can readily eliminate projection mistakes arising from occlusion and depth imprecisions while allowing high-frequency update. In Fig. 2, for example, occlusions are responsible for the repetition of image content in the vicinities of the foreground objects of the warped image. These regions, together with some segments along depth discontinuities, are identified through the low-frequency matching test.

(a) (b)

Fig. 2. (a) Ghosting artifacts for the warped view due to occlusions for the *Breakdancers* sequence, view 1, and (b) regions eliminated by the low-frequency SAD matching test (in black).

2.3. Super-resolution for multiple views

The existence of more than two views increases the amount of high- frequency content available for super-resolution and reduces the effects of occlusions and depth map errors. Without loss of generality, we consider the case of 3 views and super-resolve the central image V_n^L by transferring high-frequency content from adjacent images V_{n-1}^H and V_{n+1}^H. In this case, when content from both adjacent images is available, V_n^H is estimated through a weighted average in which weights are based on the quality of low-frequency matching, i.e, the inverse of the respective SAD's. Then,

$$V_n^H(u,v) = \frac{\frac{V_{n-1}'^H(u',v')}{\mathrm{SAD}_{n-1}^L} + \frac{V_{n+1}'^H(u',v')}{\mathrm{SAD}_{n+1}^L}}{\frac{1}{\mathrm{SAD}_{n-1}^L} + \frac{1}{\mathrm{SAD}_{n+1}^L}}, \quad (5)$$

where SAD_{n-1}^L and SAD_{n+1}^L are computed between the low-frequency image intensities V_n^L and $V_{n-1}'^L$ or $V_{n+1}'^L$, respectively. If only one adjacent view has a valid projection, (4) applies for that pixel. The super-resolved image if formed by $\hat{V}_n = V_n^L + V_n^H$.

3. EXPERIMENTAL RESULTS

The proposed super-resolution method was tested on two publicly available data sets: *Ballet* and *Breakdancers* [12]. The sequences are furnished with corresponding depth maps. Due to the limited amount of high-frequency content in the available images, we have chosen as our reference images two down-sampled versions: *Ballet* of size 512x384 and *Breakdancers* of size 256x192.

In Table 1 and 2 we present PSNR and SSIM [13] results, with respect to the reference images, obtained from up-sampling and super-resolving the low-resolution versions of view 1, *i.e.* V_1^L and $\hat{V}_1$, respectively. The low-resolution images were obtained using down-sampling factors (M) of 2 and 4. The stereo pair for these simulations is composed of views 1 and 2, and a SAD threshold of 100.0 within a 3x3 window was employed in all tests.

Sequence	**M**	**PSNR** V_1^L	**PSNR** $\hat{V}_1$
Ballet	2	38.8 dB	39.4 dB
Ballet	4	35.9 dB	37.4 dB
Breakdancers	2	39.3 dB	41.1 dB
Breakdancers	4	35.4 dB	37.7 dB

Table 1. PSNR of up-sampled V_1^L and super-resolved $\hat{V}_1$ versions with stereo pair from view 2.

Sequence	**M**	**SSIM** V_1^L	**SSIM** $\hat{V}_1$
Ballet	2	0.94	0.94
Ballet	4	0.85	0.90
Breakdancers	2	0.95	0.96
Breakdancers	4	0.85	0.91

Table 2. SSIM of up-sampled V_1^L and super-resolved $\hat{V}_1$ versions with stereo pair from view 2.

Note that for both sequences the super-resolution approach significantly boosts PSNR measures. For the *Ballet* sequence, super-resolution achieves gains of 0.6 and 1.5 dB over the up-sampled versions for M=2 and M=4, respectively. For the *Breakdancers* sequence, PSNR gains are 1.8 and 2.3 dB. The SSIM metric reflects more modest gains.

In Table 3 and 4 we present PSNR and SSIM results obtained from up-sampling (V_1^L) and super-resolving ($\hat{V}_1$) low-resolution versions of view 1 using neighboring views 0 and 2. Low-resolution images were also obtained using down-sampling factors of 2 and 4.

The use of 2 adjacent views contributes towards improvements in terms of PSNR and, to a lesser extent, SSIM. When compared to the stereo setups of Tables 1 and 2, the *Ballet*

Sequence	M	PSNR V_1^L	PSNR $\hat{V}_1$
Ballet	2	38.8 dB	39.8 dB
Ballet	4	35.9 dB	37.8 dB
Breakdancers	2	39.3 dB	41.8 dB
Breakdancers	4	35.4 dB	38.3 dB

Table 3. PSNR of up-sampled V_1^L and super-resolved $\hat{V}_1$ versions with views 0 and 2.

Sequence	M	SSIM V_1^L	SSIM $\hat{V}_1$
Ballet	2	0.94	0.95
Ballet	4	0.85	0.91
Breakdancers	2	0.95	0.97
Breakdancers	4	0.85	0.93

Table 4. SSIM of up-sampled V_1^L and super-resolved $\hat{V}_1$ versions with views 0 and 2.

sequences shows an improvement of 0.4 dB for both down-sampling factors while *Breakdancers* presents improvements of approximately 0.6 dB.

(a) (b)

Fig. 3. Details on the *Breakdancers* sequence, view 1 and M=2, for (a) up-sampled V_1^L and (b) super-resolved $\hat{V}_1$ using views 0 and 2.

Visually, the images also show improvements as can be seen in Fig. 3. High-frequency details are particularly visible in the foreground objects such as sharper contours on the dancers's faces and textures on their clothes. Note that for proper comparisons results should be viewed on a screen.

4. CONCLUSIONS

This paper presents a super-resolution method for use in a mixed resolution multiview framework. The low-resolution image is up-sampled to its original dimensions and super-resolved with high-frequency information from adjacent views, based on the correspondences indicated by the associated depth maps. Results indicate a significant gain in PSNR over up-sampled, non-resolved images, as well as gains of the general multiview case with respect to the stereo scenario.

A major hurdle in our tests was to find multiview sequences allying high-frequency details with reliable depth maps. The popular sequences we obtained lack high-frequency detail and hinder super-resolution tests, which forced us to down-sample them in order to serve as meaningful test subjects. We hope to soon enhance our test set with other suitable sequences.

Future work includes the investigation of alternative correspondence methods between views, such as mixed resolution depth maps and depth estimation from low-resolution images, as well as the effects of these methods on the super-resolution process. The performance of super-resolution will also be investigated under simulcast and multiview coding scenarios.

5. REFERENCES

[1] Y. Chen, M. Hannuksela, L. Zhu, A. Hallapuro, M. Gabbouj, H. Li, "Coding techniques in multiview video coding and joint multiview video model," *Picture Coding Symposium*, pp.1-4, 6-8 May 2009.

[2] B. Julesz, *Foundations of cyclopean perception*, University of Chicago Press, 1971.

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

[4] H. Brust, A. Smolic, K. Mueller, G. Tech, T. Wiegand, "Mixed resolution coding of stereoscopic video for mobile devices," *3DTV Conference*, pp.1-4, 2009.

[5] C. Fehn, P. Kauff, S. Cho, N. Hur, J. Kim, "Asymmetric coding of stereoscopic video for transmission over T–DMB," *Proceedings of 3DTV-CON – Capture, Transmission and Display of 3D Video*, Kos Island, Greece, May 2007.

[6] E. Ekmekcioglu, S. Worrall, A. Kondoz, "Utilisation of downsampling for arbitrary views in multi–view video coding," *IEEE Electronics Letters*, v. 44, pp. 339-340, 2008.

[7] A. Aksay, C. Bilen, E. Kurutepe, T. Ozcelebi, G. Akar, R. Civanlar, A. Tekalp, "Temporal and spatial scaling for stereoscopic video compression," *14th European Signal Processing Conference*, 2006.

[8] Y. Chen, Y. Wang, M. Gabbouj, M. Hannuksela, "Regionally adaptive filtering for asymmetric stereoscopic video coding", *Proceedings of ISCAS*, pp. 2585–2588, May 2009.

[9] H. Sawhney, Y. Guo, K. Hanna, R. Kumar, S. Adkins, S. Zhou, "Hybrid stereo camera: an IBR approach for synthesis of very high resolution stereoscopic image sequences," *ACM SIGGRAPH*, pp. 451–460, 2001.

[10] F. Brandi, R. de Queiroz, D. Mukherjee, "Super-resolution of video using key frames and motion estimation," *ICIP*, pp. 321–324, Oct. 2008.

[11] P. Kauff, N. Atzpadin, C. Fehn, M. Müller, O. Schreer, A. Smolic, R. Tanger, "Depth map creation and image–based rendering for advanced 3DTV services providing interoperability and scalability", *Image Commun.*, v. 22, pp. 217–234, 2007.

[12] C. Zitnick; S. Kang; M. Uyttendaele, S. Winder, R. Szeliski, "High-quality video view interpolation using a layered representation," *ACM SIGGRAPH'04*, pp. 600-608, 2004.

[13] Z. Wang, A. Bovik, H. Sheikh, E. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *Image Processing, IEEE Transactions on*, v. 13, no.4, pp.600-612, April 2004.