Spatiotemporal Super-Resolution for Multiview Video in Transform Domain

Yawei Li [#1], Xiaofeng Li [#2], Zhizhong Fu [#3], Tingting Niu [#4], Keyu Long [*5]

[#] *School of Communication and Information Engineering*
University of Electronic Science and Technology of China, Chengdu 611731, China
[1,4] {lywuestc, tinyniu1990}@163.com, [2,3] {xfli, fuzz}@uestc.edu.cn

[*] *The Second Research Institute of CAAC, Chengdu 610041, China*
[5] 51389610@qq.com

***Abstract*—Super-resolution (SR) algorithms for multiview videos aim at refining the low-resolution (LR) views using high-frequency detail information from adjacent high-resolution (HR) views. However, to refine a LR view, the state-of-the-art methods only exploit the spatial redundancy of a multiview video. To take full advantage of the spatial and temporal redundancy, a new spatiotemporal SR algorithm is proposed in this paper. Depth-image-based rendering (DIBR) establishes coordinate correspondences between different views. Motion estimation handles object motion between different frames. The high-frequency details are extracted in transform domain and added to LR images. In order to merge the details from different HR images, a combination of global and local weights is proposed. The effectiveness of the proposed algorithm is proved theoretically and shown experimentally.**

***Index Terms*—Spatial redundancy, temporal redundancy, super-resolution, multiview video, discrete cosine transform**

I. Introduction

Recently, multiview video processing has attracted considerable attention of researchers. Due to the huge cost and bandwidth requirements to store and transmit multiview video, the multiview video plus depth (MVD) format is exploited as an efficient representation method. In order to further reduce the data size, the mixed resolution (MR) format is proposed. The MR-MVD representation allows some views of a video sequence to be decimated and stored in low resolution (LR) while the others to be stored in the original high resolution (HR) [1]–[4]. Afterwards, the display system utilizes super-resolution (SR) algorithms to recover the LR views.

Depth image based rendering (DIBR) is a basic technique used to establish coordinate correspondences between different views and to synthesize the missing high-frequency detail of the LR views [5]. Garcia *et al.* proposed a SR algorithm that took advantage of the high-frequency information in the HR views to increase the visual quality of the LR views [1]. The authors used sum of absolute difference as a threshold to reject the mismatched coordinate correspondences. The same authors introduced consistency check to tackle the mismatched coordinates [2]. Since the SR algorithm in [2] is sensitive to depth error, Richter *et al.* proposed a robust SR algorithm that used displacement estimation and compensation to make up for pixel displacements due to depth error. Besides exploiting the spatial redundancy of MR-MVD video, one can also utilize temporal redundancy to extract high-frequency information. Brandi *et al.* used high-frequency information from the key frames to refine the non-key frames [6]. Similarly, Ancuti *et al.* used HR still photographs to recover the missing details of a LR frame [7]. Lengyel *et al.* devised a non-local means (NLM) algorithm to reconstruct the high-frequency content of the LR frames [8].

Compared with spatial-domain operation, high-frequency information extraction and combination in discrete cosine transform (DCT) domain, also have a compelling prospect. Hung *et al.* reported the advantages of the transform-domain SR algorithm [9]. Fu *et al.* designed an algorithm that used all available HR views to extract detail information. The extracted details were then merged with respect to the optimal weights determined by the correlation between a LR view and the adjacent HR views [10]. However, the lack of using temporal redundancy prohibited the SR algorithm from advancing.

In this paper, we propose a spatiotemporal SR algorithm that takes full advantage of the spatial and temporal redundancies of a MR-MVD video. DIBR or the so-called view projection projects the adjacent HR views onto the LR view. Motion estimation is introduced to make up for the motion between consecutive frames. A combination of the global image weight and the local block weight is proposed to merge the details extracted from the HR views. The rest of the paper is organized as follows. Section 2 describes the multiview setup of the proposed algorithm. Section 3 explains the SR algorithm in detail. Section 4 proves the merits of using both spatial and temporal redundancies. Section 5 shows the experimental results. Section 6 concludes the paper.

II. The Multiview Setup

The multiview setup is shown in Fig. 1, where View V_κ is the LR view to be refined, View $V_{\kappa-1}$ and $V_{\kappa+1}$ are the HR views. The notations $\tau-1$, τ, and $\tau+1$ represent three consecutive time indices. We refer to the images $I_{\kappa-1,\tau}$ and $I_{\kappa+1,\tau}$ as the spatial neighbors of $I_{\kappa,\tau}$ while $I_{s,t}(s=\kappa-1,\kappa+1;t=\tau-1,\tau+1)$ as the spatiotemporal neighbors. For the simplicity of notation, all the neighbors

978-1-5090-5316-2/16/$31.00

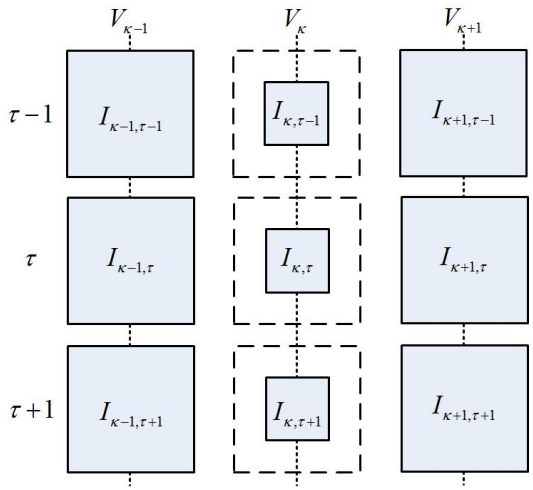


Fig. 1. Multiview setup.

of $I_{\kappa,\tau}$ are denoted as $I_n(n = 1, \cdots, N)$. The aim is to enhance the visual quality of $I_{\kappa,\tau}$ with the aid of all its neighbors. The setups for other LR images are identical. In this paper, we exploit both view projection and motion estimation to compensate for the displacements between $I_{\kappa,\tau}$ and its neighbors. View projection is implemented between the HR view and LR view. In order to further compensate for the motion between different frames, motion estimation is carried out after view projection.

III. The Proposed Algorithm

Fig. 2 shows the flowchart of the proposed algorithm. The inputs consist of HR images and LR images. The input images are preprocessed for the sequel operation. We first interpolate the LR image $I_{\kappa,\tau}$ to the same resolution as HR images, resulting in its low-frequency part $I^L_{\kappa,\tau}$. The HR images I_n undergo blurring, decimation, and interpolation to generate their low-frequency parts I^L_n. Meanwhile, the HR images are warped to the LR view via view projection.

I^L_n and $I^L_{\kappa,\tau}$ are used for motion estimation and the calculation of global and local weights. Blockwise motion estimation is enforced between $I^L_{\kappa,\tau}$ and its spatiotemporal neighbors. Global weights are calculated between $I^L_{\kappa,\tau}$ and its neighbors while local weights are calculated for the blocks of images. For every target LR block, the high-frequency details are extracted from HR images and are merged to form the desired high-frequency information. At last, the result is added to the low-frequency parts of the LR frames.

A. View Projection

View projection is used to establish coordinate correspondences between different views. We assume the well-known pinhole camera model. The pixel (u, v) in $I_{\kappa+1,\tau}$ is first projected to the 3D position, yielding

$$(x, y, z)^T = R^{-1}_{\kappa+1} A^{-1}_{\kappa+1} D(u, v)(u, v, 1)^T + t_{\kappa+1} \quad (1)$$

where $R_{\kappa+1}$, $A_{\kappa+1}$, and $t_{\kappa+1}$ represents the rotation matrix, the intrinsic matrix, and the translation vector of the camera in

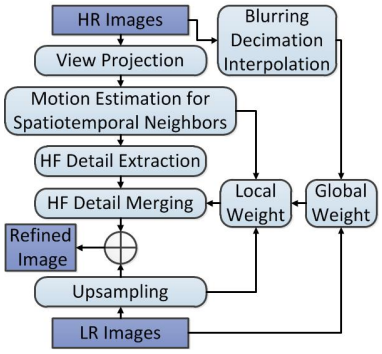


Fig. 2. Flowchart of the proposed algorithm.

view $V_{\kappa+1}$, respectively. $D(u, v)$ is the physical depth value of pixel (u, v). Then the 3D point is projected to the image plane of $I^L_{\kappa,\tau}$, i.e.

$$w(u', v', 1)^T = A_\kappa R_\kappa [(x, y, z)^T - t_\kappa]. \quad (2)$$

B. Motion Estimation

Since there is relative motion between $I^L_{\kappa,\tau}$ and its spatiotemporal neighbors even after view projection, we need to implement motion estimation to compensate for it. We define $I^L_{s,t}(s = \kappa - 1, \kappa + 1; t = \tau - 1, \tau + 1)$ as the target frames and $I^L_{\kappa,\tau}$ as the anchor frame. Let B_p be a block in the anchor frame. The aim is to find a matching block B_q in every target frame that minimizes the mean square error, i.e.

$$B_q = \arg\min \mathcal{E}_r = \arg\min \|B_p(\mathbf{X}) - B_i(\mathbf{X} + \mathbf{D})\|^2_2 \quad (3)$$

where $\mathbf{X}$ is the position of block B_p, $\mathbf{D}$ is the relative motion between the B_p and B_i. We uses the exhaustive block-matching algorithm (EBMA) to find the best match for each block in the target frame.

C. Global Weight Calculation

Different views of multiview videos have varying degrees of correlation depending on the spatial distance between them [10]. The optimal weight calculation in [10] doesn't satisfy real-time display because it needs the whole video sequence. Actually, we find that the frames with the same time indices between different views have almost the same degree of similarity. Meanwhile, we also need to evaluate the similarity of different frames in the same view. Therefore, we define the global weight of I^L_n as the correlation coefficient between $I^L_{\kappa,\tau}$ and I^L_n, i.e.

$$\omega^g_n = \frac{\sum\sum (I^L_{\kappa,\tau} - \bar{I}^L_{\kappa,\tau})(I^L_n - \bar{I}^L_n)}{\sqrt{\left(\sum\sum (I^L_{\kappa,\tau} - \bar{I}^L_{\kappa,\tau})^2\right)\left(\sum\sum (I^L_n - \bar{I}^L_n)^2\right)}} \quad (4)$$

where $\bar{I}^L_{\kappa,\tau}$ and $\bar{I}^L_n$ are the means of $I^L_{\kappa,\tau}$ and I^L_n.

D. Local Weight Calculation

The local weight is calculated blockwise. Different blocks of two images have varying degrees of similarities. Thus, it's useful to incorporate the local weight into the weighting

procedure. Let B^I be the interpolated block of a LR block B in $I_{\kappa,\tau}$. After view projection and motion estimation, B^I has a matching block B_n in every I_n^L. B_n^H is the HR counterpart of B_n in the image I_n. Let the mean square error between B^I and B_n be $\mathcal{E}_n = \left\|B^I - B_n\right\|_2^2$. Then the local weight of B_n is

$$\omega_n^l = \exp(-\mathcal{E}_n / 2\sigma_1^2) \tag{5}$$

where σ_1 is the decaying factor. Having obtained the global and local weights, we multiply them and normalize them to achieve the final complete weights, i.e.

$$\omega_n = \omega_n^g \omega_n^l / \sum_n \omega_n^g \omega_n^l \tag{6}$$

E. High-Frequency Component Extraction and Merging

High-frequency information extraction and merging is carried out in DCT domain. We transform the HR blocks B_n^H to DCT domain and represent the derived DCT coefficients by a blocked matrix,

$$\mathcal{B}_n = \begin{bmatrix} \mathcal{B}_n^{00} & \mathcal{B}_n^{01} \\ \mathcal{B}_n^{10} & \mathcal{B}_n^{11} \end{bmatrix} \tag{7}$$

where $\mathcal{B}_n^{01}$, $\mathcal{B}_n^{10}$, and $\mathcal{B}_n^{11}$ are the high-frequency components, $\mathcal{B}_n^{00}$ is an inaccurate low-frequency component. The derived high-frequency components from multiple HR images are summed with respect to the weights ω_n and constitute the detail information of B. The DCT coefficients of B acts as the low-frequency components of its refined block. Thus, we have the merged DCT coefficients

$$\hat{\mathcal{B}} = \begin{bmatrix} 2\mathcal{B} & \sum_n \omega_n \mathcal{B}_n^{01} \\ \sum_n \omega_n \mathcal{B}_n^{10} & \sum_n \omega_n \mathcal{B}_n^{11} \end{bmatrix} \tag{8}$$

At last, $\hat{\mathcal{B}}$ is transformed to spatial domain, resulting in the super-resolved block of B.

IV. Weighted Sum of High Frequency Detail

In this section, we prove the validness of exploiting both spatial and temporal redundancies of MR-MVD videos. The captured pixel value at position $\mathbf{x}$ is the actual light intensity corrupted by additive white Gaussian noise (AWGN), i.e.

$$\psi(\mathbf{x}) = \varphi(\mathbf{x}) + \delta(\mathbf{x}) \tag{9}$$

where $\psi(\mathbf{x})$, $\varphi(\mathbf{x})$ and $\delta(\mathbf{x})$ represent the captured pixel value, the light intensity, and Gaussian noise, respectively. The Gaussian noise is with zero mean and variance σ^2.

With view projection and motion estimation, a series of measurements of the same position can be observed,

$$\psi_n(\mathbf{x}) = \varphi(\mathbf{x}) + \delta_n(\mathbf{x}), n = 1, 2, \cdots, N \tag{10}$$

where $\delta_n(\mathbf{x})$ are independent identically distributed (i.i.d.) random variables. The mean and variance of $\psi_n(\mathbf{x})$ are $\varphi(\mathbf{x})$ and σ^2, respectively. Here, we compute the weighted sum of the observed measurements as

$$\hat{\psi}(\mathbf{x}) = \sum_{n=1}^{N} \omega_n \psi_n(\mathbf{x}) = \varphi(\mathbf{x}) + \sum_{n=1}^{N} \omega_n \delta_n(\mathbf{x}) \tag{11}$$

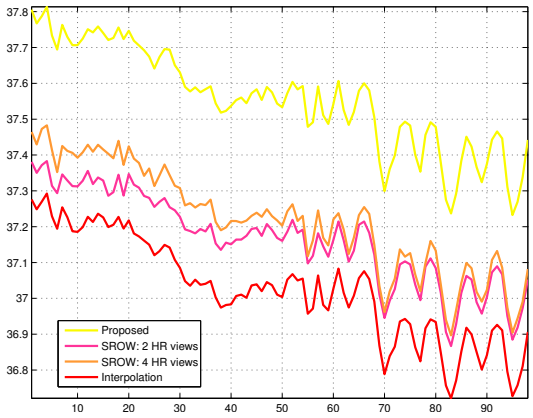


Fig. 3. PNSR results for *Ballet*.

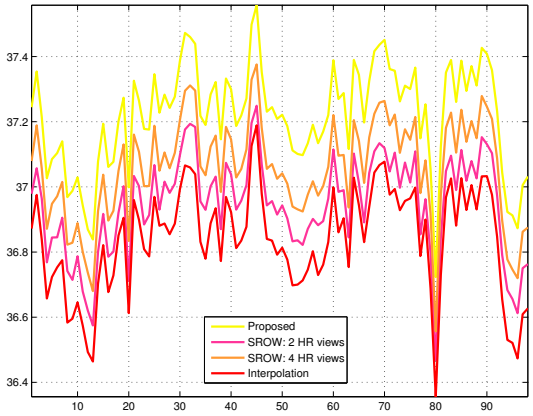


Fig. 4. PSNR results for *Breakdancers*.

where ω_n are the computed weights subject to $\sum \omega_n = 1$. Since $\delta_n(\mathbf{x})$ are i.i.d random variables, the mean and variance of the computed statistic are

$$E\left[\hat{\psi}(\mathbf{x})\right] = \varphi(\mathbf{x}) \tag{12a}$$

$$Var\left[\hat{\psi}(\mathbf{x})\right] = \left(\sum_{n=1}^{N} \omega_n^2\right) \sigma^2 \tag{12b}$$

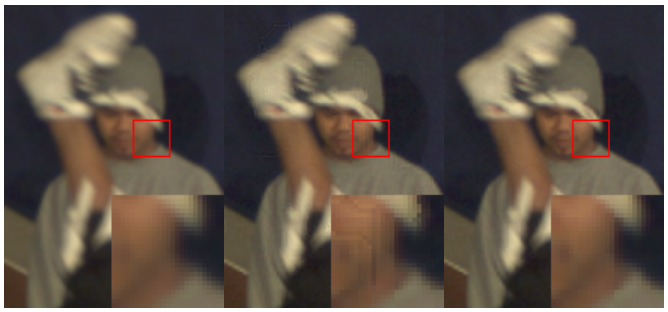
Considering the constrains of the weights, we have the result that $\sum \omega_n^2 < 1$. In conclusion, the weighted sum $\hat{\psi}(\mathbf{x})$ has the same mean as $\psi_n(\mathbf{x})$ but smaller variance. Thus, the noise is reduced and $\hat{\psi}(\mathbf{x})$ has better performance than $\psi_n(\mathbf{x})$.

V. Experimental Results

In this section, we shows the experimental results. We test our SR algorithm on the sequences *Ballet* and *Breakdancers* [11].The size of the blocks for motion estimation and high-frequency component extraction is 8×8. The search area for motion estimation is 31×31. The decaying factor is set to 5. We adopt a 2-tap two-dimensional Lanczos filter to blur the HR frames and downsample the result with the downsampling rate being 2. For the the proposed algorithm, View 3 is selected as the LR view while View 2 and View 4 are the HR views.

(a) (b) (c)

Fig. 5. (a) - (c) are details of bilinear interpolation, SROW, and the proposed method for *Ballet*.



(a) (b) (c)

Fig. 6. (a) - (c) are details of bilinear interpolation, SROW, and the proposed method for *Breakdancers*.

Peak Signal-to-Noise Ratio (PSNR) and Structural SIMilarity (SSIM) [12] are used to compare the performance of different SR algorithms.

We compare the proposed algorithm with SR using optimal weights (SROW) [10] and SR using sum of absolute difference (SRSAD) [1]. Two or four HR views are available for SROW while one or two views are used for SRSAD. In contrast, only two views are available for the proposed algorithm. Since the spatiotemporal neighbors are considered, six images are actually used to refine a LR image. The SR results of Frame 65 are listed in Table 1. SRSAD has very limited improvements. When two HR views are available for SROW, the proposed algorithm outperforms it with a large margin. When the available HR views are mounted to four, the proposed approach still works better. Fig. 3 and Fig. 4 depict more detailed results of the experiment for all the LR frames in View 3. It's clear that our method leads to higher PSNR results under all the circumstances.

Fig. 5 and Fig. 6 show the details of the refined images. Both SROW and the proposed method refine the image with high-frequency content, compared with interpolation. However, SROW results in blurring artifacts while our method generates better results.

TABLE I
THE PSNR AND SSIM RESULTS FOR FRAME 65 OF BOTH *Ballet* AND *Breakdancers*

Method	*Ballet* PSNR/SSIM	*Breakdancers* PSNR/SSIM
Interpolated	37.0565/0.9862	36.9435/0.9863
SRSAD: 1 HR view	37.0629/0.9862	36.9471/0.9863
SRSAD: 2 HR views	37.0685/0.9862	36.9502/0.9863
SROW: 2 HR views	37.2055/0.9890	37.0121/0.9844
SROW: 4 HR views	37.2324/0.9891	37.1409/0.9886
Proposed	**37.5793/0.9893**	**37.3436/0.9888**

VI. CONCLUSION

In this paper, we propose a novel SR algorithm that uses both spatially and temporally adjacent images to refine a LR image of a MR video. The core idea of this approach is to take advantage of the high-frequency information usually ignored by other methods in the spatiotemporal frames. In order to merge the computed high-frequency information efficiently, the global weight based on correlation of the whole images and the local weight based on mean square error of blocks are introduced. The frequency decomposition and integration are carried out in DCT domain. The experimental results show that both the subjective and objective SR performance are improved by our SR method.

REFERENCES

[1] D. C. Garcia, C. Dorea, and R. L. de Queiroz, "Super-resolution for multiview images using depth information," in *Proc. ICIP*, 2010, pp. 1793 – 1796.

[2] ——, "Super resolution for multiview images using depth information," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 9, pp. 1249 – 1256, 2012.

[3] T. Richter and A. Kaup, "Multiview super-resolution using high-frequency synthesis in case of low-framerate depth information," in *Proc. VCIP*, 2012, pp. 1 – 6.

[4] T. Richter, J. Seiler, W. Schnurrer, and A. Kaup, "Robust super-resolution for mixed-resolution multiview image plus depth data," *IEEE Trans. Circuits Syst. Video Technol.*, 2015.

[5] C. Fehn, "Depth-image-based rendering (DIBR), compression and transmission for a new approach on 3D-TV," *Proc. SPIE*, vol. 5291, pp. 93 – 104, 2004.

[6] F. Brandi, R. d. Queiroz, and D. Mukherjee, "Super-resolution of video using key frames and motion estimation," in *Proc. ICIP*, 2010, pp. 321 – 324.

[7] C. Ancuti, C. O. Ancuti, and P. Bekaert, "Video super-resolution using high quality photographs," in *Proc. ICASSP*, 2010, pp. 862 – 865.

[8] R. Lengyel, S. M. Reza Soroushmehr, and S. Shirani, "Multi-view video super-resolution for hybrid cameras using modified NLM and adaptive thresholding," in *Proc. ICIP*, 2014, pp. 5437 – 5441.

[9] E. M. Hung, C. Dorea, D. C. Garcia, and R. L. de Queiroz, "Transform-domain super-resolution for multiview images using depth information," in *Proc. EUSIPCO*, 2011, pp. 398 – 401.

[10] Z. Fu, Y. Li, J. Xu, H. Wu, and Y. Lai, "Super resolution for multiview mixed resolution images in transform-domain with optimal weight," *Multimed. Tools Appl.*, pp. 1–15, 2016.

[11] C. L. Zitnick, S. B. Kang, M. Uyttendaele, S. Winder, and R. Szeliski, "High-quality video view interpolation using a layered representation," *SIGGRAPH*, vol. 23, no. 3, pp. 600 – 608, 2004.

[12] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600 – 612, 2004.