Manifold Learning Based Super Resolution for Mixed-Resolution Multi-view Video in Visual Internet of Things

Yuan Zhou[1,3(✉)], Ying Wang[2], Yeda Zhang[3], Xiaoting Du[3], Hui Liu[1], and Chuo Li[3]

[1] The National Ocean Technology Center, Tianjin 300111, China
zhouyuan@tju.edu.cn
[2] Unit 61660 of PLA, Beijing 100089, China
[3] Tianjin University, Tianjin 300072, China

Abstract. In a Visual Internet of Things (VIoT), the video sequences of different viewpoints are captured by different visual sensors and transmitted simultaneously, which puts a huge burden on storage and bandwidth resources. Mixed-resolution multi-view video format can alleviate the burden on the limited storage and bandwidth resources. However, the low resolution view need to be up-sampled to provide high quality visual experiences to the users. Therefore, a super resolution (SR) algorithm to reconstruct the low resolution view is highly desirable. In this paper, we propose a new two-stage super resolution method. In the first depth-assisted high frequency synthesis stage, depth image based rendering (DIBR) is used to project a high resolution view to a low resolution view to estimate the super resolution result. Then in the second high frequency compensation stage, the local block matching model based on manifold learning is used to enhance the super resolution result. The experimental results demonstrate that our method is capable to achieving a PSNR gain up to 4.76 dB over bicubic baseline and recover details in edge regions, without sacrificing the quality of smooth areas.

Keywords: Visual Internet of Things · Super resolution · Manifold learning

1 Introduction

In recent years, the field of Visual Internet of Things (VIoT) has attracted a lot of research attention and provided a broad variety of application, such as security surveillance, smart homes, health-care [5,14]. VIoT is comprised of a large number of visual sensors, each of which captures massive video information

Supported by organization NSFC61571326.

S. Han et al. (Eds.): AICON 2019, LNICST 287, pp. 486–495, 2019.
https://doi.org/10.1007/978-3-030-22971-9_41

about target events, and transmits them to a central base station or a data sink. However, the video sequences of different viewpoints are captured by different visual sensors and transmitted simultaneously in a common VIoT, which puts a huge burden on storage and bandwidth resources. In order to reduce the amount of data to be compressed, the mixed-resolution (MR) multi-view video format is introduced [13]. In the MR video format, some visual sensors acquire viewpoints at high resolution (HR), while others are captured at low resolution (LR). This significantly reduces the amount of data for storage and transmission for multi-view video applications.

Recently, Brust *et al.* propose mixed resolution multi-view coding framework for mobile devices, where one of the views is coded entirely at a lower spatial resolution than the others [3].The multi-view video plus depth (MVD) representation and coding scheme are proposed in [1,11], which provide a multi-view texture scene and the corresponding depth map. These methods can provide fairly good results in bitrate reduction. However, the decoded low-resolution view in these aforementioned work is of poor visual quality compared with the high resolution view, since these methods do not compensate for the quality differences between the views.

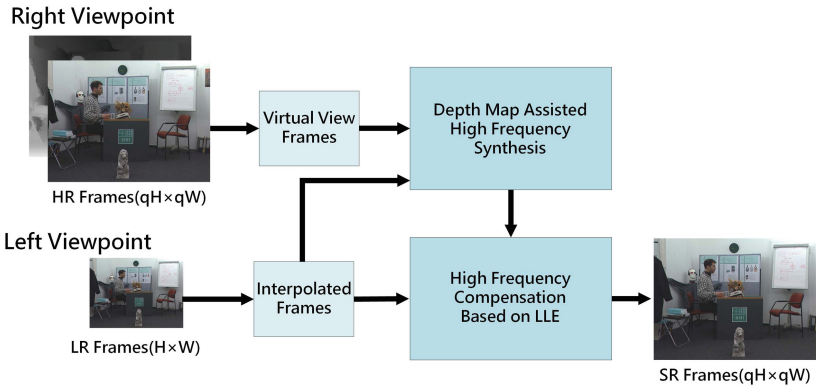

Fig. 1. The framework of our proposed method. The high resolution right view assists the SR of low resolution left view. The SR method contains two stages: (1) the depth map assisted high frequency component synthesis (DHFS) stage; and (2) the high frequency compensation (HFCOM) stage.

We consider situations where two views are of different spatial resolutions. According to binocular suppression theory [2], the human visual system (HVS) perceives high frequency information from the high resolution view, such that one can approximately perceive a visual feeling of the high resolution view. However, such approximate feeling will cause visual uncomfortableness. In order to tackle this deficiency, a super resolution (SR) algorithm to reconstruct the low resolution view is highly desirable. In general, super resolution algorithms can be categorized into three classes, namely the interpolation-based [18],

reconstruction-based [16] and example-based approaches [4,15]. In recent years, these SR resolution methods have been applied to multi-view video applications gradually. Besides, for multi-view video SR, high frequency information from neighboring high-resolution views can be used to increase the visual quality of the low-resolution camera video sequence [9,10,17].

In this paper, we propose a novel multi-view video SR method assisted by virtual views. Unlike the aforementioned papers intending to obtain more accurate high frequency information from adjacent high resolution views, we present an example learning based method to compensate the synthesized high resolution information. Figure 1 depicts the framework of our proposed method, which consists of two stages: (1) the depth map assisted high frequency component synthesis (DHFS) stage; and (2) the high frequency compensation (HFCOM) stage. To better recover high frequency information, we mainly contribute to the multi-view video SR in the following two aspects:

(1) In the first stage, we measure the similarity of the virtual and interpolated views using both the spatial and depth information to locate mismatched regions, and reduce erroneous high frequency added to the interpolated view; and

(2) In the second stage, a block matching method based on locally linear embedding (LLE), a representative algorithm in manifold learning, is used to compensate the synthesized high frequency component and reconstruct the missing high frequency component in the mismatched regions, which is able to remarkably improve the quality of reconstruction, especially for the non-parallel multi-view video sequences.

2 The Depth Assisted High Frequency Synthesis

In the original high frequency synthesis method [8], the high frequency component is directly extracted from the virtual view $\boldsymbol{V_L^v}$ and added to the interpolated view $\boldsymbol{V_L^i}$. However, this virtual view generated using the depth-image-based rendering (DIBR) [6] method contains correspondence errors such as cracks and occlusions attributable to the inherent defects of DIBR. As a result, the high frequency component extracted from the correspondence errors will cause obvious mismatching in the SR result. Therefore, a similarity check between the virtual view $\boldsymbol{V_L^v}$ and the interpolated view $\boldsymbol{V_L^i}$ is employed to identify mismatched regions, and the high frequency part of the projection points in $\boldsymbol{V_L^v}$ that fail in the similarity check would not be added to the corresponding position of $\boldsymbol{V_L^i}$.

Similarity is measured by calculating the sum of the absolute differences (SAD) of corresponding image blocks centered around $\boldsymbol{(u', v')}$ between the virtual view $\boldsymbol{V_L^v}$ and the interpolated view $\boldsymbol{V_L^i}$. To further utilize the depth information, we also calculate SAD of the corresponding depth map blocks centered around $\boldsymbol{(u', v')}$ between the virtual depth map $\boldsymbol{D_L^v}$ and the original depth map $\boldsymbol{D_L}$. The total SAD of the corresponding blocks in the color images and depth maps is shown as follows

$$SAD = \sum_{(u',\, v') \in S} \left| V_L^v\,(u',\ v') - V_L^i\,(u',\ v') \right| + \sum_{(u',\, v') \in S} \left| D_L\,(u',\ v') - D_L^v\,(u',\ v') \right| \tag{1}$$

where $\boldsymbol{S}$ is the block range, and $D_L^v\,(u',\ v')$ is the projected virtual depth map, which is mentioned in the last section.

The occlusion and crack areas in virtual view will have obvious difference with the interpolated view, and the depth value of these areas in the virtual depth map would also have difference with the original depth map, these would lead to a large SAD value. If the SAD value of block $\boldsymbol{S}$ is larger than a threshold $\boldsymbol{T_s}$, all the pixels in block $\boldsymbol{S}$ of $\boldsymbol{V_L^v}$ are refilled by the interpolated pixels from $\boldsymbol{V_L^i}$, and we mark $\boldsymbol{M\,(u',\ v') = 0}$, or all the pixels in block $\boldsymbol{S}$ remain to be virtual view pixels and $\boldsymbol{M\,(u',\ v') = 1}$. Here, $\boldsymbol{M}$ is a mask matrix that indicates whether or not a pixel is from the virtual view.

After the similarity check, the virtual view result can be refined. One can extract the low frequency component of $\boldsymbol{V_L^v}$ using a Gaussian filter $\boldsymbol{F}$ and add the high frequency component, which is the difference of the original virtual view frame and its low frequency component, to the interpolated view to obtain the (DHFS) result $\boldsymbol{V_L^{SYN}}$ as follows

$$V_L^{SYN} = V_L^i + (V_L^v - F(V_L^v)) \times M. \tag{2}$$

3 High Frequency Compensation Based on Manifold Learning

In our high frequency compensation model, we use the local linear embedding (LLE) [12] algorithm, which is a representative algorithm in manifold learning, to enhance the high frequency synthesized in the first stage. The DHFS super resolution result $\boldsymbol{V_L^{SYN}}$ is assumed as the HR space, while its down sample denoted by $\boldsymbol{V_{L_d}^{SYN}}$ is assumed as the LR space. Unlike using the whole training set as the searching range, we determine a much smaller searching range centering around the LR image patch. Thus, block matching is processed between the low-resolution frame $\boldsymbol{V_L}$ and $\boldsymbol{V_{L_d}^{SYN}}$ in a small searching range, and then a fusion of the high frequency information extracted from the matched blocks is added to the interpolated image to enhance the SR result.

Figure 2 illustrates the manifold learning based high frequency compensation model, where $\boldsymbol{x_n}$ indicates the image patches in the low-resolution frame $\boldsymbol{V_L}$, while $\boldsymbol{y_n}$ refers to the SR result corresponding to $\boldsymbol{x_n}$. Here, $\boldsymbol{y_n = y_n^I + y_n^H}$, where $\boldsymbol{y_n^I}$ denotes the interpolation result of $\boldsymbol{y_n}$ and $\boldsymbol{y_n^H}$ is the high frequency part reconstructed by the LLE algorithm. The high frequency compensation (HFCOM) algorithm follows three steps:

(1) Determine the local searching range. To enhance the current patch $\boldsymbol{x_n}$ in $\boldsymbol{V_L}$, a corresponding local neighbor area $\boldsymbol{R_n}$ in $\boldsymbol{V_{L_d}^{SYN}}$ is determined as

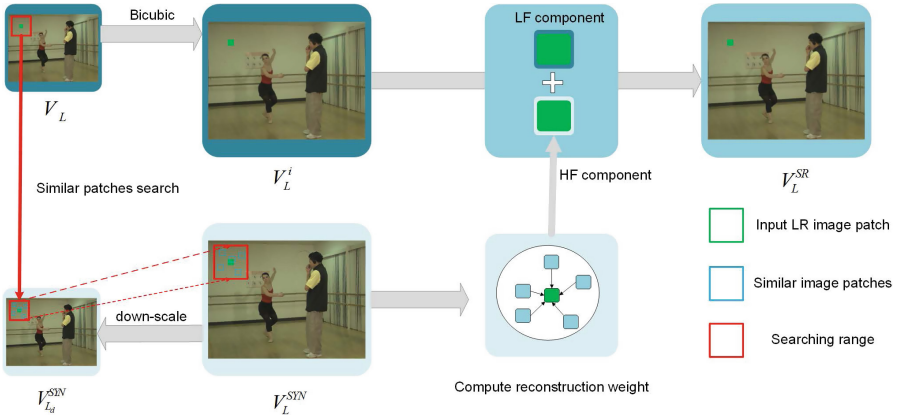

Fig. 2. Manifold learning based high frequency compensation model.

the searching area. The size of the searching area is determined by whether or not $\boldsymbol{x_n}$ is a high-frequency missing patch. A high-frequency missing patch is determined to the ratio of the number of interpolated pixels to the whole patch. For each patch $\boldsymbol{x_n}$, we calculate

$$\lambda_{x_n} = 1 - \frac{\sum_{(i,j)\in R_n} M(i,j)}{S' \times S'} \tag{3}$$

where, $\boldsymbol{S'}$ is the size of block $\boldsymbol{x_n}$. $\boldsymbol{M}$ is the mask map calculated in last section. If $\boldsymbol{\lambda_{x_n}}$ is larger than threshold $\boldsymbol{\lambda_0}$, $\boldsymbol{x_n}$ is determined as a high-frequency missing patch, and the search range of $\boldsymbol{x_n}$ will be the whole image space of $\boldsymbol{V_{L_d}^{SYN}}$. Otherwise, the search range will be a small square area that centers around $\boldsymbol{x_n}$.

(2) Search similar image patches and obtain reconstruction weights. Then several similar patches of $\boldsymbol{x_n}$ are searched in $\boldsymbol{R_n}$ and indicated as $\boldsymbol{x_n^k}$. The reconstruction weights $\boldsymbol{W_n}$ are computed by minimizing the local reconstruction error.

The Euclidean distance is used to define neighbors and reconstruct weights of $\boldsymbol{x_n}$. Several similar patches, which have the smallest Euclidean distance with $\boldsymbol{x_n}$, are found and marked as $\boldsymbol{x_n}, k = 1, \ldots, K$. By minimizing the local reconstruction error for $\boldsymbol{x_n}$, the optimality is achieved:

$$\varepsilon_n = \left\| x_n - \sum_{k=1}^{K} w_k x_n^k \right\| \quad s.t. \sum_{k=1}^{K} w_k = 1. \tag{4}$$

Apparently, minimizing $\boldsymbol{\varepsilon_n}$ subject to the constraints is a constrained least square problem. Define a local Gram matrix

$$G_n \triangleq (x_n I^T - X_n)^T (x_n I^T - X_n) \tag{5}$$

where $\boldsymbol{I}$ is a column vector of ones and $\boldsymbol{X_n}$ is a matrix with its columns being the neighbors of $\boldsymbol{x_n}$. An efficient way to solving (10) is to solve $\boldsymbol{G_n W_n} = \boldsymbol{I}$, and then normalize the weights. Then, the reconstruction weights $\boldsymbol{W_n}$ of patch $\boldsymbol{x_n}$ is attainted.

(3) Reconstruct the high frequency component. These similar patches have corresponding image patches in $\boldsymbol{V_L^{SYN}}$, which are indicated as $\boldsymbol{y_n^k}$. Finally, by fusing the high frequency part of $\boldsymbol{y_n^k}$ under the same reconstruction weights $\boldsymbol{W_n}$, the high frequency part $\boldsymbol{y_n^H}$ is reconstructed.

The high-frequency part of $\boldsymbol{x_n}$ is reconstructed based on the corresponding HR patches with the aid of weigh matrix $\boldsymbol{W_n}$ and the high-frequency parts of $\boldsymbol{y_n^k}$ as shown

$$y_n^H = \sum_{k=1}^{K} w_k \left(y_n^k - F\left(y_n^k \right) \right) \tag{6}$$

where $\boldsymbol{F}$ is the Gaussian used to obtain the low frequency component. And the reconstructed patch is

$$\boldsymbol{y_n} = \boldsymbol{y_n^I} + \boldsymbol{y_n^H} \tag{7}$$

After repeating the above steps for all the patches in $\boldsymbol{V_L}$, all the reconstructed patches are obtained. In order to enforce the inter-patch relationships and to weaken the block effects, the pixels in overlapping region of the reconstructed image are averaged.

4 Experimental Results

4.1 Parameters Setting

We test our SR method on typical multi-view video sequences, including both parallel and non-parallel camera video sequences, such as *Ballet* and *Breakdancer*. The low resolution video sequence is generated by downsampling one of the HR view sequences. A $\mathbf{5 \times 5}$ Gaussian kernel is been used to blur the original high-resolution view prior to down-sampling. The down-sampling factor is two for both the horizontal and vertical directions.

4.2 Experiment Result on Multi-view Video

Table 1 shows the PSNR and SSIM values of the final super-resolution results obtained by the proposed method, bicubic interpolation, the original HFSYN [7], and VVA [10]. The PSNR and SSIM values were calculated over several the frames of each test video and then averaged. It can be seen that the proposed method is able to significantly boost the PSNR performance. For instance, the proposed method achieves average gains of 4.76 dB over the bicubic interpolation method, 4.10dB over [7] and 2.84 dB over [10]. As for SSIM, it achieves average gains of 0.0304 over the bicubic interpolation method, 0.0309 and 0.0203 over [7] and [10], respectively.

Table 1. PSNR (dB) and SSIM comparative results of different methods.

Dataset	Bicubic	HFSYN [7]	VVA [10]	Proposed
BookArrial	32.44/0.9101	33.92/0.9256	35.11/0.9451	**37.16/0.9580**
DoorFlowers	32.81/0.9262	34.05/0.9375	35.68/0.9459	**37.72/0.9615**
LeavingLaptop	32.77/0.9094	34.04/0.9253	35.15/0.9273	**37.63/0.9578**
Champtower	33.06/0.9656	33.80/0.9614	35.50/0.9688	**39.39/0.9854**
Pantomime	32.54/0.9598	34.17/0.9634	34.85/0.9721	**39.46/0.9869**
Ballet	33.79/0.9231	33.61/0.8962	33.95/0.9205	**37.36/0.9433**
Breakdancer	36.95/0.9064	35.38/0.8876	36.30/0.8909	**38.96/0.9206**
Average	33.48/0.9286	34.14/0.9281	35.40/0.9387	**38.24/0.9590**

(a) Bicubic (33.87 / 0.9231)

(b) HFSYN (33.66 / 0.8960)

(c) VVA (34.04 / 0.9207)

(d) Proposed (**37.37** / **0.9431**)

Fig. 3. *Ballet* SR result. The visual results and PSNR(dB)/SSIM values of different methods. Our method recover the finest strip pattern.

Figures 3 and 4 show sample frames from the video sequences *Ballet* and *Doorflowers* with a up-scaling factor of two. The Bicubic method is used as the baseline. Figures 3(b) and 4(b) are the original high frequency synthesis (HFSYN) result. The HFSYN method [7] is capable of restoring image details,

(a) Bicubic (32.83 / 0.9263)

(b) HFSYN (34.08 / 0.9378)

(c) VVA (35.67 / 0.9460)

(d) Proposed (**37.63** / **0.9615**)

Fig. 4. *Doorflowers* SR result. The visual results and PSNR(dB)/SSIM values of different methods. Our method reconstructs the clearest boundary.

albeit with evident artifacts along the boundary between the object and the background. These artifacts are caused by mismatched high frequency components. Figures 3(c) and 4(c) are virtual view assisted SR (VVA) result. The VVA method [10] can process a much better visual quality than HFSYN result, but still suffers from some undesirable stair-like artifacts and distortion. The results produced by our proposed method as presented in Figs. 3(d) and 4(d) can reconstruct sharper edges and yield better structure details. Our method can effectively tackle the mismatch problem. Neither stair-like nor substantial blurring artifacts occur in our recovering method, showing much better visual results than other comparative methods.

5 Conclusion

In this paper, we propose a novel two-stage method for efficient mixed-resolution multiview super resolution based on refined virtual view synthesis and manifold learning. The first depth map assisted high frequency component synthesis stage can effectively reduce mismatching regions in the synthesis virtual view.

The second high frequency compensation stage can reconstruct the missing high frequency in the mismatching regions and enhance the visual quality. Experiment shows that our method can get remarkable gain in both quantitative and qualitative result.

References

1. Aflaki, P., et al.: Coding of mixed-resolution multiview video in 3D video application. In: 2013 20th IEEE International Conference on Image Processing (ICIP), pp. 1704–1708. IEEE (2013)
2. Blake, R.: Threshold conditions for binocular rivalry. J. Exp. Psychol. Hum. Percept. Perform. **3**(2), 251 (1977)
3. Brust, H., Smolic, A., Mueller, K., Tech, G., Wiegand, T.: Mixed resolution coding of stereoscopic video for mobile devices. In: 3DTV Conference, pp. 1–4. Citeseer (2009)
4. Chang, H., Yeung, D.Y., Xiong, Y.: Super-resolution through neighbor embedding. In: Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition CVPR 2004, p. I (2004)
5. Cruz, M.A.A.D., Rodrigues, J.J.P.C., Al-Muhtadi, J., Korotaev, V., Albuquerque, V.H.C.: A reference model for internet of things middleware. IEEE Internet Things J. **99**, 1 (2018)
6. Fehn, C.: Depth-image-based rendering (DIBR), compression, and transmission for a new approach on 3D-TV. In: Stereoscopic Displays and Virtual Reality Systems XI, vol. 5291, pp. 93–105. International Society for Optics and Photonics (2004)
7. Garcia, D.C., Dorea, C., de Queiroz, R.L.: Super resolution for multiview images using depth information. IEEE Trans. Circuits Syst. Video Technol. **22**(9), 1249–1256 (2012)
8. Garcia, D.C., Drea, C., Queiroz, R.L.D.: Super-resolution for multiview images using depth information. In: IEEE International Conference on Image Processing, pp. 1793–1796 (2010)
9. Jain, A.K., Nguyen, T.Q.: Video super-resolution for mixed resolution stereo. In: 2013 20th IEEE International Conference on Image Processing (ICIP), pp. 962–966. IEEE (2013)
10. Jin, Z., Tillo, T., Yao, C., Xiao, J., Zhao, Y.: Virtual-view-assisted video super-resolution and enhancement. IEEE Trans. Circuits Syst. Video Technol. **26**(3), 467–478 (2016)
11. Merkle, P., Smolic, A., Muller, K., Wiegand, T.: Multi-view video plus depth representation and coding. In: IEEE International Conference on Image Processing. ICIP 2007, vol. 1, pp. 1–201. IEEE (2007)
12. Roweis, S.T., Saul, L.K.: Nonlinear dimensionality reduction by locally linear embedding. Science **290**(5500), 2323 (2000)
13. Sawhney, H.S., Guo, Y., Hanna, K., Kumar, R., Adkins, S., Zhou, S.: Hybrid stereo camera: an IBR approach for synthesis of very high resolution stereoscopic image sequences. In: Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques. pp. 451–460. ACM (2001)
14. Sezer, O.B., Dogdu, E., Ozbayoglu, A.M.: Context aware computing, learning and big data in internet of things: a survey. IEEE Internet Things J. **5**(1), 1–27 (2017)
15. Yang, C.Y., Huang, J.B., Yang, M.H.: Exploiting self-similarities for single frame super-resolution. In: Asian Conference on Computer Vision, pp. 497–510 (2010)

16. Zhang, J., Cao, Y., Wang, Z.: A simultaneous method for 3D video super-resolution and high-quality depth estimation. In: ICIP, pp. 1346–1350 (2013)
17. Zhang, J., Cao, Y., Zha, Z.J., Zheng, Z., Chen, C.W., Wang, Z.: A unified scheme for super-resolution and depth estimation from asymmetric stereoscopic video. IEEE Trans. Circuits Syst. Video Technol. **26**(3), 479–493 (2016)
18. Zhang, X., Wu, X.: Image interpolation by adaptive 2-D autoregressive modeling and soft-decision estimation. IEEE Trans. Image Process. **17**(6), 887–896 (2008)