Multiview Video Super-Resolution via Information Extraction and Merging

Yawei Li, Xiaofeng Li, Zhizhong Fu, Wenli Zhong
School of Communication and Information Engineering
University of Electronic Science and Technology of China, Chengdu, China
lywuestc@163.com, {xfli, fuzz}@uestc.edu.cn, zwluestc@163.com

ABSTRACT

Multiview video super-resolution provides a promising solution to the contradiction between the huge data size of multiview video and the degraded video quality due to mixed-resolution compression. This algorithm consists of two different functional layers. An information extraction layer draws relevant high-frequency information from the high-resolution views via depth-image-based rendering and interpolation. A merging layer fuses multiview high-frequency information to refine the low-resolution view. In this paper, we introduce kernel regression and non-local means to improve the two layers, respectively. Kernel regression adapts to the local image structure and thus outperforms basic interpolation methods. Non-local means exploits the similarity between different views of multiview videos to restore the high-frequency component of a low-resolution image. We constrain non-local means by limiting the pixels used to restore a pixel. The experimental results show the effectiveness of the proposed algorithm.

Keywords

Super-resolution; kernel regression; non-local means; multiview video processing

1. INTRODUCTION

Super-resolution (SR) algorithms aim at recovering the high-frequency component of a low-resolution (LR) image and creating its high-resolution (HR) counterpart. Since the rediscovery by Huang *et al.* [9], SR algorithms have shown a promising prospect in medical image processing, video compression, three-dimensional television (3DTV), and free-viewpoint television (FTV). Compared with expensive hardware capture system, SR algorithms provide an economic alternative to generate HR images and videos.

SR algorithms are divided into reconstruction-based methods and example-based methods [13]. Reconstruction-based methods use subpixel displacement among successive LR images. Goldlueche *et al.* first proposed a 3D multiview SR approach that led to a total variation energy for the texture map [7]. By solving the Euler-Lagrange equation on the surface, the texture map could be recovered. Example-based methods establish a library from a set of training samples, which guides the learning phase of the SR algorithm [4].

When applied to a video sequence, SR algorithms exploit the temporal or spatial redundancy of the video instead of establishing libraries from irrelevant training samples. Brandi *et al.* proposed an algorithm that super-resolved non-key frames of a video sequence with the assisted information from key frames [2]. Ancuti *et al.* used HR still photographs of the same scene to refine a LR video [1]. In addition, mixed-resolution (MR) multiview video allows one to exploit the spatial redundancy of videos to increase the sharpness of a LR view [6]. MR video coding is proposed to save the cost and bandwidth used to capture, transmit, and store multiview videos. However, this video format fails to offer satisfactory viewing experience in different perspectives due to the inherent LR views. Fortunately, multiview video SR algorithms solve the problem [6, 12].

Multiview video SR methods mainly contain an information extraction layer and an information merging layer. The extraction layer is based on depth-image-based rendering (DIBR) that projects high-frequency content from a HR view to a LR view [3, 6]. On the other hand, the merging layer is a mechanism that fuses all of the available high-frequency information to recover a LR view. Since the algorithm proposed by Garcia *et al.* resulted in projection errors in case of noisy depth map [6], Richter *et al.* proposed an algorithm that was based on displacement-compensated high frequency synthesis and aimed at correcting the errors [12]. Besides the algorithms extracting information in spatial domain, Fu *et al.* extracted high-frequency content in discrete cosine transform (DCT) domain and devised an optimal weight distribution to merge multiview information [5]. Jin *et al.* proposed a decision mechanism to determine whether to fill the missing pixels in the super-resolved frame by HR-view pixels or by interpolated LR-view pixels [10].

In this paper, we propose to use kernel regression and non-local means (NLM) to improve the two functional layers of multiview SR algorithms, respectively. Kernel regression is based on the assumption of the local continuity of natural images and changes the shape of the kernel according to the local image structure [8]. Thus, it can avoid introducing irrelevant pixels into predicting an unknown pixel and improves the performance of the extraction layer. Due to the occlusion of objects and the displacement of cameras, the derived images by the extraction layer usually contain holes

MM '16, October 15-19, 2016, Amsterdam, Netherlands

DOI: http://dx.doi.org/10.1145/2964284.2967260

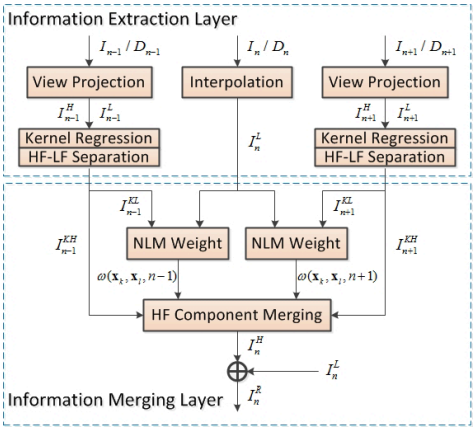


Figure 1: Flowchart of the proposed algorithm.

and cracks that cannot be filled by a single view. Thus, the information merging layer is needed to fuse the multiview images. Contrary to the assumption of local continuity of kernel regression, NLM assumes that a similar patch can repeat within a non-local area of an image [11]. We impose constraints on the number of available pixels to reject the dissimilar ones which may be misused to recover a pixel.

The rest of the paper is organized as follows. Section 2 describes the proposed algorithm in detail. Section 3 shows the experimental results. Section 4 concludes the paper.

2. THE PROPOSED ALGORITHM

Multiview video plus depth (MVD) format is used for the SR algorithm. Let I_n be the LR image, I_{n-1} and I_{n+1} be the HR images in the adjacent view. In addition, each view is provided with a depth map D (See Fig. 1). The video format is not limited to just two HR views and can be extended to any possible HR configurations. Since the operations of I_{n-1} and I_{n+1} are symmetric, we mainly talk about I_{n-1}.

The proposed algorithm contains two layers: an information extraction layer and a merging layer. In the information extraction layer, the LR image is enlarged by bicubic interpolation, resulting in its full-resolution counterpart I_n^L. However, I_n^L lacks high-frequency information and needs to be enhanced with the aid of I_{n-1} and I_{n+1}. Thus, the HR images undergo view projection, kernel regression. View projection establishes coordinate correspondences between I_n^L and I_{n-1} and projects high-frequency information from I_{n-1} to the image plane of I_n^L. As is shown in Fig. 2, the derived coordinates are irregular samples (blue stars) instead of the regular samples (red points). Thus, kernel regression is used to restore the regular samples. Meanwhile, component separation is implemented along with kernel regression, resulting in a high-frequency component I_{n-1}^{KH} and a low-frequency component I_{n-1}^{KL}.

The primitively refined images by kernel regression have some unavoidable holes and cracks. Therefore, the merging layer is used to combine the multiview images and to alleviate the cracks and holes. The high-frequency component I_{n-1}^{KH} and the low-frequency component I_{n-1}^{KL} are used differently in NLM, which is detailed in Subsection 2.2. After NLM, the derived detail information and structure information are added to complete the whole algorithm.

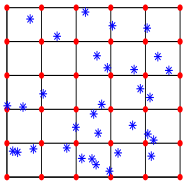


Figure 2: Regular samples (red points) and irregular samples (blue stars).

2.1 The Information Extraction Layer

2.1.1 *View Projection*

View projection is a fundamental technology in MR-MVD processing. It establishes coordinate correspondences between a LR view and its adjacent HR view.

We assume a basic pinhole model where points in space and points in an image plane are linked via central projection. To start with, a point (x_1', x_2') in the HR image plane of I_{n-1} is projected to its space position

$$(u_1, u_2, u_3)^T = R_{n-1}^{-1} A_{n-1}^{-1} D(x_1', x_2')(x_1', x_2', 1)^T + t_{n-1} \quad (1)$$

where A_{n-1}, R_{n-1}, and t_{n-1} are the intrinsic matrix, rotation matrix, and translation vector of the camera, respectively. $D(x_1', x_2')$ is the physical depth which can be derived from the depth map, i.e.

$$\frac{1}{D(x_1', x_2')} = \frac{d(x_1', x_2')}{255}\left(\frac{1}{D_{min}} - \frac{1}{D_{max}}\right) + \frac{1}{D_{max}} \quad (2)$$

where $d(x_1', x_2')$ is the 8-bit entry in the depth map, D_{max} and D_{min} define the physical depth range. Then the space point is projected onto the image plane of I_n^L, i.e.

$$w(x_1, x_2, 1)^T = A_n R_n \left[(u_1, u_2, u_3)^T - t_n\right] \quad (3)$$

Therefore, a point (x_1', x_2') in I_{n-1} is uniquely associated with a point (x_1, x_2) in the image plane of I_n^L. Ideally, the pixel information of (x_1', x_2') can be directly used to recover that of (x_1, x_2). However, the derived points are irregular samples(Fig. 2). In order to obtain the regular samples, different methods are used. Fu *et al.* directly rounded irregular samples to their regular positions, thus incurring at most half-pixel error along each axis [5]. Garcia *et al.* obtained regular samples via bilinear interpolation [6]. In the following subsection, we will introduce kernel regression to obtain regular samples.

2.1.2 *Kernel Regression*

Kernel regression assumes the local continuity of natural images and exploits this local regularity to restore images and videos. It's suitable not only for interpolation of regularly sampled images but also for restoration of irregularly sampled images. We discuss recovering regular samples from irregular samples here (Fig. 2). For the simplicity of notation, we use a coordinate vector $\mathbf{x}$ to denote a pixel position. Kernel regression is generalized as a least-square optimization problem, i.e.

$$\widehat{z}(\mathbf{x}_i) = \arg\min_z \sum_{\mathbf{x}_j \in \mathcal{N}(\mathbf{x}_i)} (I_{n-1}(\mathbf{x}_j) - z)^2 K_{\mathbf{x}_i}(\mathbf{x}_j - \mathbf{x}_i) \quad (4)$$

where $K_{\mathbf{x}_i}(\mathbf{x}-\mathbf{x}_i)$ is the steering kernel, $\mathcal{N}(\mathbf{x}_i)$ is the neighborhood of $\mathbf{x}_i$, $\mathbf{x}_i$ and $\mathbf{x}_j$ denote the regular and irregular samples. Solving the optimization problem gives

$$\widehat{z}(\mathbf{x}_i) = \frac{\sum\limits_{\mathbf{x}_j \in \mathcal{N}(\mathbf{x}_i)} I_{n-1}(\mathbf{x}_j) K_{\mathbf{x}_i}(\mathbf{x}_j - \mathbf{x}_i)}{\sum\limits_{\mathbf{x}_j \in \mathcal{N}(\mathbf{x}_i)} K_{\mathbf{x}_i}(\mathbf{x}_j - \mathbf{x}_i)} \tag{5}$$

Thus, the irregular neighbors $\mathbf{x}_j$ constitute the regression basis of $\mathbf{x}_i$. The essence of kernel regression is to develop a steering kernel adapted to the local structure of an image. An adaptive Gaussian kernel can be constructed, i.e.

$$K_{\mathbf{x}_i}(\mathbf{x}-\mathbf{x}_i) = \frac{\{\det(C_i)\}^{1/2}}{2\pi\sigma_k^2} \exp\left\{\frac{(\mathbf{x}-\mathbf{x}_i)^T C_i^{-1}(\mathbf{x}-\mathbf{x}_i)}{-2\sigma_k^2}\right\} \tag{6}$$

where C_i is the diffusion tensor at $\mathbf{x}_i$ controlling the shape of the kernel, σ_k controls the bandwidth of the kernel [15].

2.1.3 Component Separation

The bicubic interpolation I_n^L of the LR image offers convincing low-frequency (structure) information although it lacks high-frequency (detail) information. Instead of restore the overall pixel value, one only need to recover the detail information of the LR image I_n. Therefore, a proper separation of the detail and structure information becomes an important task for the information extraction layer.

The component separation of regular samples is implemented simultaneously with kernel regression, i.e.

$$\widehat{z}(\mathbf{x}_i) = I_{n-1}^{KL}(\mathbf{x}_i) + I_{n-1}^{KH}(\mathbf{x}_i) \tag{7}$$

where $I_{n-1}^{KL}(\mathbf{x}_i)$ and $I_{n-1}^{KH}(\mathbf{x}_i)$ are the detail and structure components of $\mathbf{x}_i$. These two components are computed as

$$I_{n-1}^{KL}(\mathbf{x}_i) = \frac{\sum\limits_{\mathbf{x}_j \in \mathcal{N}(\mathbf{x}_i)} I_{n-1}^{L}(\mathbf{x}_j) K_{\mathbf{x}_i}(\mathbf{x}_j - \mathbf{x}_i)}{\sum\limits_{\mathbf{x}_j \in \mathcal{N}(\mathbf{x}_i)} K_{\mathbf{x}_i}(\mathbf{x}_j - \mathbf{x}_i)} \tag{8a}$$

$$I_{n-1}^{KH}(\mathbf{x}_i) = \frac{\sum\limits_{\mathbf{x}_j \in \mathcal{N}(\mathbf{x}_i)} I_{n-1}^{H}(\mathbf{x}_j) K_{\mathbf{x}_i}(\mathbf{x}_j - \mathbf{x}_i)}{\sum\limits_{\mathbf{x}_j \in \mathcal{N}(\mathbf{x}_i)} K_{\mathbf{x}_i}(\mathbf{x}_j - \mathbf{x}_i)} \tag{8b}$$

where I_{n-1}^L and I_{n-1}^H are the low-frequency and high-frequency parts of the HR image I_{n-1}. I_{n-1}^L is derived from I_{n-1} by blurring, decimation, and interpolation. I_{n-1}^H is the difference between I_{n-1} and I_{n-1}^L [2]. In the information extraction layer, the low-frequency component guides the fusion of multiview high-frequency components.

2.2 The Information Merging Layer

2.2.1 Image Merging

The images derived from kernel regression have annoying artifacts of holes and cracks. For example, Fig. 3 shows this phenomenon in two complementary views. Each of the views has artifacts that cannot be filled with information from a single HR view. However, if the two images in Fig. 3 are merged, the resultant image will overcome this problem. Since I_n^L has convincing low-frequency information, we only need to merge multiview high-frequency information via NLM.

In MR multivew setup, NLM exploits the similarity between a LR image and its adjacent HR views to recover

(a) (b)

Figure 3: (a) and (b) are two complementary views.

the missing high-frequency information of I_n^L. For a pixel in I_n^L, its high-frequency component is computed as a weighted sum of those of its neighbors in the HR images, i.e.

$$I_n^H(\mathbf{x}_k) = \frac{\sum\limits_{v} \sum\limits_{\mathbf{x}_l \in \mathcal{S}(\mathbf{x}_k, v)} I_v^{KH}(\mathbf{x}_l)\omega(\mathbf{x}_k, \mathbf{x}_l, v)}{\sum\limits_{v} \sum\limits_{\mathbf{x}_l \in \mathcal{S}(\mathbf{x}_k, v)} \omega(\mathbf{x}_k, \mathbf{x}_l, v)} \tag{9}$$

where $\mathcal{S}(\mathbf{x}_k, v)$ is the similarity patch that contains the neighbors of pixel $\mathbf{x}_k$ in the HR view v, $\mathbf{x}_l$ is a pixel in the patch, and $\omega(\mathbf{x}_k, \mathbf{x}_l, v)$ is a weight that reflects the similarity of $\mathbf{x}_k$ and $\mathbf{x}_l$. The weight is computed by comparing the two low-frequency parts, i.e.

$$\omega(\mathbf{x}_k, \mathbf{x}_l, v) = \exp\left(\frac{-\left\|(R_{\mathbf{x}_k} I_v^{KH} - R_{\mathbf{x}_l} I_n^L) G_\sigma\right\|_2^2}{2\sigma_l^2}\right) \tag{10}$$

where the operator $R_{\mathbf{x}}$ extracts a patch around $\mathbf{x}$, σ_l is a smoothing factor, G_σ is a Gaussian kernel with standard deviation σ.

2.2.2 Constrained NLM

One major problem of the described algorithm is that it involves too many pixels in the recovery of a pixel in the LR image. Although a dissimilar pixel can be assigned with a small weight, it still participates in the weighted-averaging process. In the extreme case, considerable dissimilar pixels that are used to recover a pixel will accumulate restoration errors. Thus, we choose no more than ρ pixels in the weighting process to reject the dissimilar pixels. The selected ρ pixels are those with the largest possible weights.

2.3 Discussion

From (5) and (9), kernel regression and NLM have similar computing structures. However, they act differently in the method. First, the basic assumption of the two methods are different. Kernel regression uses local continuity and smoothness to interpolate regular samples while NLM exploits the similarity of multiview video. Second, as mentioned above, they play different roles in the proposed method.

Some other work is also related to combining the advantages of kernel regression and NLM [15, 16]. However, the multiview setup makes the discussed algorithm distinct. The MR format allows the algorithm to learn detail information

Table 1: Parameter Setup

Parameter	σ_k	σ_l	σ	ρ	$\mathcal{N}(\mathbf{x})$ size	$\mathcal{S}(\mathbf{x})$ size
Value	2.6	1.6	1	12	3	7

Table 2: Comparison of Intermediate Results and the Final Result

Method	PSNR(dB)	SSIM
Bilinear Interpolation	36.6643	0.9882
No separtion, no constraint	36.6854	0.9709
Only Constraint	37.1654	0.9806
Final Result	**38.7299**	**0.9925**

Table 3: PSNR (dB) and SSIM of Frame 20

Method	Ballet PSNR/SSIM	Breakdancers PSNR/SSIM
Bicubic	37.0085/ 0.9890	36.8104/0.9883
SRTD	37.3724/ 0.9893	37.1304/0.9884
SRDM	38.5339/ 0.9927	38.0414/0.9894
Proposed	**39.0107/ 0.9933**	**38.8205/0.9925**

from the HR view although no learning library is trained. View projection establishes relative accurate coordinate correspondences, which makes it possible for kernel regression to compute regular samples. Component separation guarantees accurate high-frequency information is extracted.

3. EXPERIMENTAL RESULTS

In this section, we validate the proposed approach by experimental results. We use the sequences *Ballet* and *Breakdancers* for our tests [17]. The parameter setups are given in Table 1. For all of the SR algorithms, View 3 is the LR view to be refined with the aid of HR View 0, View 2, View 4, and View 6. In order to simulate the LR view, the original full-resolution images in View 3 are down-sampled with a sampling rate of 2.

We conduct two experiments with different purposes. The first experiment compares the intermediate results and the final result of the proposed algorithm to show how each technique improves the method's performance. The results in comparison are bicubic interpolation, result without component separation and constrained NLM (No separtion, no constraint), result only with constrained NLM (Only Constraint), and the final result. The Peak Sigal-to-Noise Rate (PSNR) and Structure SIMilarity (SSIM) [14] of *Breakdancers* are shown in Table 2.

In the second experiment, we compare the proposed algorithm with bicubic interpolation, SR in transform domain (SRTD) [5], and SR for multiview images using depth map (SRDM) [6]. Table 3 shows the PSNR and SSIM results of Frame 20 of the two sequences. SRTD can only provide limited improvement over bicubic interpolation. The PSNR gains are only 0.3639dB and 0.33dB for *Ballet* and *Breakdancers*, respectively. Compared with SRTD, SRDM has relative better result. Among the SR algorithms, the propsed has the best performance. The PSNR gain of the proposed algorithm over the other methods are 2.0022dB, 1.6381 dB and 0.4768 dB for *Ballet*. Fig. 4 shows the PSNR result of 30 frames of *Ballet*. The proposed algorithm has the best performance under all the circumstances.

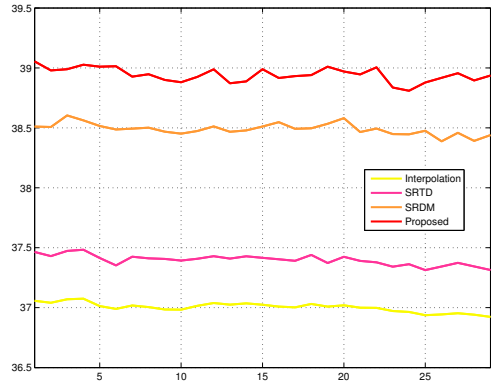
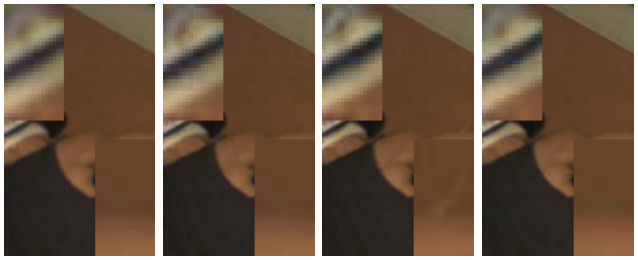
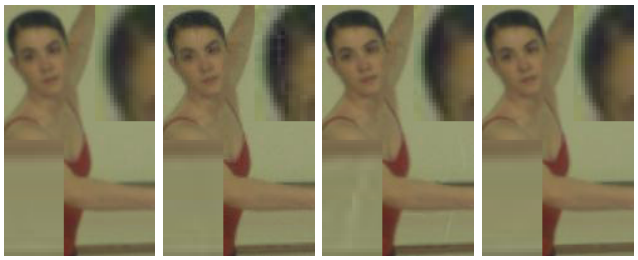


Figure 4: PSNR (dB) of 30 frames for *Ballet*



(a) (b) (c) (d)

Figure 5: (a) - (d) are the details of bicubic interpolation, SRTD, SRDM, and the proposed method for *Breakdancers*.



(a) (b) (c) (d)

Figure 6: (a) - (d) are the details of bicubic interpolation, SRTD, SRDM, and the proposed method for *Ballet*.

Fig. 5 and Fig. 6 shows the details of the SR images. Both SRDM and the proposed algorithm can recover satisfactory high-frequency information for the LR image. However, SRDM has some artifacts around the edges of the images.

4. CONCLUSION

In this paper, we proposed a multiview SR algorithm that improves the performance of the information extraction layer and the merging layer. Kernel regression adapted to the local structure extracts more accurate high-frequency information, compared with basic interpolation. NLM is used to merge the extracted high-frequency information from HR views. The constraints imposed on the number of available pixels avoid incorporating mismatched pixels into the restoration of a pixel. The experimental results show the effectiveness of the main techniques used in the proposed algorithm. The comparison with other multiview SR algorithms shows the improvement of the proposed algorithm.

5. REFERENCES

[1] C. Ancuti, C. O. Ancuti, and P. Bekaert. Video super-resolution using high quality photographs. In *Proc. ICASSP*, pages 862 – 865, 2010.

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

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

[4] W. T. Freeman, T. R. Jones, and E. C. Pasztor. Example-based super-resolution. *IEEE Comput. Graph. Appl.*, 22(2):56 – 65, 2001.

[5] 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.*, pages 1–15, 2016.

[6] D. C. Garcia, C. Dorea, and R. L. de Queiroz. Super resolution for multiview images using depth information. *IEEE Trans. Circuits Syst. Video Technol.*, 22(9):1249 – 1256, 2012.

[7] B. Goldluecke and D. Cremers. Superresolution texture maps for multiview reconstruction. In *Proc. CVPR*, pages 1677 – 1684, 2009.

[8] T. Hiroyuki, F. Sina, and M. Peyman. Kernel regression for image processing and reconstruction. *IEEE Trans. Image Process.*, 16(2):349 – 366, 2007.

[9] T. S. Huang and R. Y. Tsai. Multi-frame image restoration and registration. *Adv. Comput. Vis. Image Process.*, 1:317 – 339, 1984.

[10] Z. Jin, T. Tillo, C. Yao, and J. Xiao. Virtual view assisted video super-resolution and enhancement. *IEEE Trans. Circuits Syst. Video Technol.*, PP(99):1, 2015.

[11] P. Matan, E. Michael, T. Hiroyuki, and M. Peyman. Generalizing the nonlocal-means to super-resolution reconstruction. *IEEE Trans. Image Process.*, 18(1):36 – 51, 2009.

[12] 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.*, PP(99):1, 2015.

[13] H. Su, L. Tang, Y. Wu, D. Tretter, and J. Zhou. Spatially adaptive block-based super-resolution. *IEEE Trans. Image Process.*, 21(3):1031 – 1045, 2012.

[14] 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.*, 13(4):600 – 612, 2004.

[15] H. Zhang, J. Yang, Y. Zhang, and T. S. Huang. Image and video restorations via nonlocal kernel regression. *IEEE Trans. Cybern.*, 43(3):1035 – 1046, 2013.

[16] K. Zhang, X. Gao, D. Tao, and X. Li. Single image super-resolution with non-local means and steering kernel regression. *IEEE Trans. Image Process.*, 21(11):4544 – 4556, 2012.

[17] C. L. Zitnick, S. B. Kang, M. Uyttendaele, S. Winder, and R. Szeliski. High-quality video view interpolation using a layered representation. *SIGGRAPH*, 23(3):600 – 608, 2004.