PATCH-BASED MULTIPLE VIEW IMAGE DENOISING WITH OCCLUSION HANDLING

Shiwei Zhou, Yu Hen Hu and Hongrui Jiang

University of Wisconsin, Madison, WI 53706, USA

ABSTRACT

A novel patch-based multi-view image denoising algorithm is proposed. This method leverages the 3D focus image stacks structure to exploit self-similarity and image redundancy inherent in multiple view images. Then a depth-guided adaptive window and dynamic view selection criterion is developed to aid proper selection of most consistent patches for the multi-view image denoising. Extensive experiments have been performed. Comparing the outcomes against those of state of the art image denoising algorithms, our proposed algorithm demonstrates significant performance advantage.

Index Terms— multi-view, denoising, focus image stacks, occlusion handling, non-local means

1. INTRODUCTION

Multi-view image denoising has received growing attention due to its wide application in 3D reconstruction, motion analysis, and video surveillance, etc. With more images participating in denoising, both intra-view and inter-view similarities can be exploited, promising a superior denoising quality than single-image denoising methods.

State of art single-image denoising methods such as the canonical non-local means (NLM) [2] and block matching 3D (BM3D) [3] have demonstrated good denoising quality. Based on these results, several multi-view image denoising algorithms have been proposed: Zhang et al. [1] proposed a depth-guided similarity measure for grouping patches and collaboratively denoising the patches using PCA or tensor analysis. Similarly, Luo et al. [4] proposed to apply the depth-guided joint-view distance to multiple view NLM that can adaptively select the optimal number of patches in denoising process. Meanwhile, Xue et al. [5] presented a graphical model of surface patches for patch clustering, and then applying Wiener filtering in the transformed domain to attenuate the noise. These multi-view image denoising methods achieved improved denoising quality at the expense of computational expensive exhaustive patch matching operations. Recently, Miyata et al. [6] proposed to use plane sweeping [9] for image reconstruction to perform effective multi-view image denoising with low computation cost. In [7], we introduced the 3D focus image stacks to further improve denoising quality without significantly increasing computational cost.

In this work, a new image denoising algorithm incorporating a novel occlusion handling scheme is proposed. This method extends the traditional NLM patch-based denoising procedure to multi-view images to exploit image redundancy across neighboring views. Moreover, leveraging the 3D focus image stacks, we developed novel adaptive window and view selection criteria to exclude outlier image patches from occluded views. The resulting denoised image quality shows significant improvement over all existing approaches.

In section 2, the proposed patch-based multi-view image denoising method will be presented. Experiment results comparing against existing denoising methods are presented in section 3, and we conclude the paper in section 4.

2. MULTI-VIEW IMAGE DENOISING ALGORITHM

2.1. Problem Formulation

Given a set of noisy images $\{I_{s,t}(x, y), s, t \in Z^2\}$ captured using a camera array that consists of identical cameras on a plane. Each camera's position will be inferred by a grid point (s, t), and (x, y) is the image pixel coordinate. The optical axis of each camera is perpendicular to the camera plane. We assume the noisy images is the sum of a noiseless true image and additive noise:

$$I_{s,t}(x,y) = I'_{s,t}(x,y) + n_{s,t}(x,y), \quad (1)$$

where $I'_{s,t}$ is the noiseless true image, and $n_{s,t}$ is i.i.d. zero-mean Gaussian noise with variance σ^2, i.e. $n_{s,t}(x, y) \sim N(0, \sigma^2)$. Given the target view (s_0, t_0), the objective of multi-view image denoising is to estimate $I'_{s0,t0}(x, y)$, i.e. true target view image given the set of noisy multi-view images.

2.2. Focus Image Stacks and Disparity Estimation

In [7], we proposed 3D focus image stacks as a powerful multi-view image representation to facilitate fast disparity estimation and preliminary multi-view image denoising.

Given a set of multi-view images $I_{s,t}(x, y)$ with target view located at (s_0, t_0), the images from other cameras can be aligned against the target view with respect to a fixed disparity value d by translational shifting the amount d:

$$I^{d}_{s,t}(x,y) = I_{s,t}(x+(s-s_0)d, y+(t-t_0)d). \quad (2)$$

For each disparity value d, the translated images $I^{d}_{s,t}(x, y)$ are stacked into a 3D image stack $F_d(x, y)$, which we call the *3D focus image stack*. One of the important properties of focus image stacks is that if d is correct disparity for pixel (x, y), then all the pixels at location (x, y) in the focus image stack F_d correspond to the same surface point in the 3D scene, and thus should have similar intensity values [7]. A patch-wise implementation of this property has been proposed [7] where a similarity measure S_d is computed for each disparity d:

$$S_d(x,y) = \frac{1}{N}\sum_{k=1}^{N}\sum_{(i,j)\in W(x,y)} \left| F_d(i,j,k) - I_{s_0,t_0}(i,j) \right|, \quad (3)$$

where N is the total number of cameras, $F_d(i, j, k)$ is the k^{th} slice in the focus image stack F_d, and $W(x, y)$ is an $L\times L$ window centered at (x, y). The disparity value for pixel (x, y) is then found as the d that minimizes $S_d(x, y)$. That is,

$$\hat{d}(x,y) = \arg\min_{d} S_d(x,y). \quad (4)$$

2.3. Depth-guided Adaptive Window Selection

With the 3D focus image stacks and disparity map estimated, one may estimate intensity value for each pixel by weighted averaging each pixel stack in F_d and therefore reduce noise variance [7]. While this approach has the advantage of relative low computational complexity, it falls short in denoising performance however. Firstly, due to depth variation, not all cameras in the camera array can observe a particular point on the surface of an object. The view of some camera may be occluded by other part of the object surface. The pixel values of such occluded views should be excluded from estimation. Secondly, earlier efforts [6, 7] considered only individual pixels, without taking into account neighboring pixel values, failing to leverage the spatial correlation to improve denoising performance.

Occlusion handling has been addressed in multi-view stereo research, but rarely mentioned in denoising literature. Occlusion occurs often along object boundaries where drastic depth variation is more likely. Inspired by the occlusion handling techniques in multi-view stereo by Kang et al. [8], we propose a depth-guided adaptive window selection and dynamic views selection scheme that is able to automatically select the similar patches from neighboring views while discarding occluded patches. Fig. 1 illustrates its fundamental idea using a simple three-view image system, with the middle view being the target view. Region A, B, C denote the occluding foreground, the partially occluded region, and the background, respectively. The white pixel 1 is a pixel located in region A near the boundary of A and B, while the black pixel 2 is another pixel in region B and is partially occluded by region A in the right view. As the top row of Fig. 1 illustrates, if the window centered at these pixels (the solid-line windows) is selected, regular patch matching and denoising algorithms tend to make mistakes due to the dissimilarities between these patches. By selecting the window shifting to the top-left corner, as shown in the

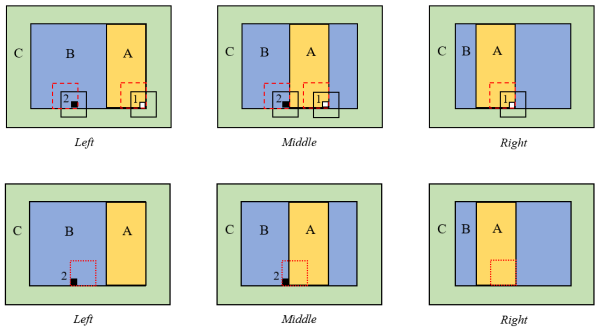
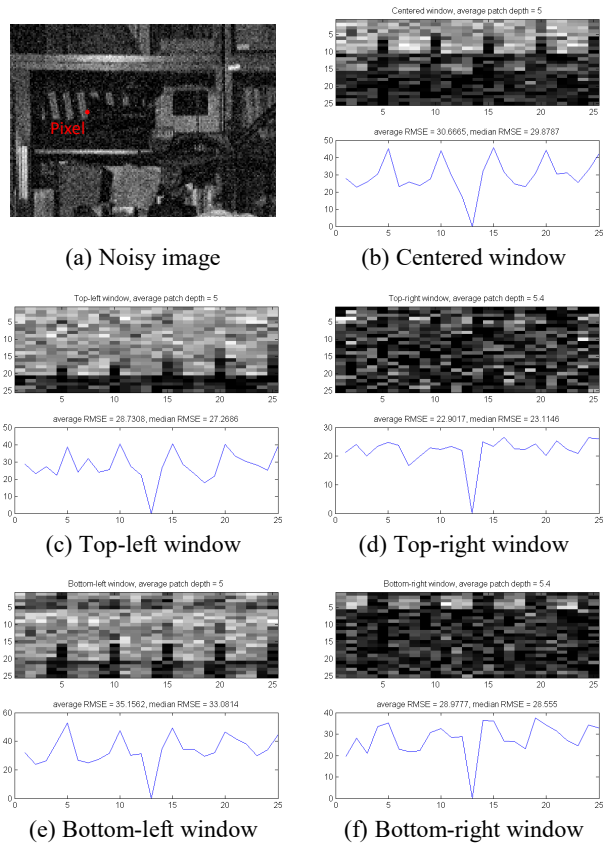


Fig. 1. Illustration of adaptive windows. The top row shows the conventional centered window, and the more reasonable top-left window. The bottom row shows the incorrect top-right window for pixel 2.

dashed-line windows, both pixels now have their patches correctly matched, and further denoising procedures can be properly applied on these patches. Note that pixel 2 might still be occluded in the right view, which will be handled with appropriate views selection to be discussed later.

In real implementation, as the images are contaminated with noise, the patch similarity may not be accurately computed, which could further bias the selection of window. As shown in the bottom row of Fig. 1, for pixel 2, the top-right window (dotted-line window), instead of the top-left window, is used for computing similarities. Assume all three views are used, then the patch similarities actually might be comparable with using the top-left window or even better, due to the corruption of noise. Since the top-right window covers a large amount of region A in the middle and right view, traditional denoising algorithms such as [2, 4, 6, 7] tend to put higher weight on the right patch when estimating the target pixel value, leading to a biased estimation.

To deal with this problem, we propose to use depth (disparity) as a guide to assist window selection. In specific, among all the candidate windows, the average disparity value within each window of the target view is first computed and compared with the target pixel disparity. Windows with closer disparity values are picked for the second round of selection. Next, the entire patch volume is selected from the focus image stack using the picked candidate windows, and root mean squared error (RMSE) between each patch and the target patch (from center view) is computed. To avoid error caused by significantly biased outliers, we choose to use the median value of RMSE instead the mean as a measure for patch similarity, and the window with lowest median value is the one to be selected. Fig. 2 shows an example of our depth-guided adaptive window selection. Five candidate windows are implemented for window selection. Without depth information as a guide, the top-right window which has the lowest RMSE median value would have been selected, causing a biased estimation of the target pixel. The patch average disparity value helps eliminate the top-right and bottom-right window first, and select the window with lowest RMSE median among the remaining windows, which all have an average disparity of 5.



(a) Noisy image (b) Centered window

(c) Top-left window (d) Top-right window

(e) Bottom-left window (f) Bottom-right window

Fig. 2. Depth-guided adaptive window selection. (a) shows the noisy image with target pixel indicated by red dot. (b)-(f) show the patch volume in vector form for each candidate window and the RMSE plot. The target view is assumed to be the center view, i.e. the middle column.

2.4. Dynamic Views Selection

With a reasonable window selected, one can extract the whole patch volume from the focus image stack for each pixel location. Previous denoising strategies simply uses the entire volume as an input for the denoising procedure, which could lead to inferior denoising performance due to patch dissimilarities. As shown in Fig. 2, although the top-left window is selected, part of the views remain inconsistent with the target patch. The dynamic view selection scheme is thus proposed to help select appropriate views such that further denoising algorithms can be performed without involving inconsistent patches.

Normally, a semi-occluded region in the target view will only be occluded in either the preceding or the succeeding views in the focus image stacks. As shown in Fig. 1, with the top-left window used for pixel 2, the patch is only occluded in the right view, while still being visible in the center and left views. The same phenomenon can also be observed in Fig. 2, where for the top-left window in (c), the 5th, 10th, 15th, 20th, and 25th views (columns) are obviously contaminated by the occluding object (the video recorder) which has lower intensity values.

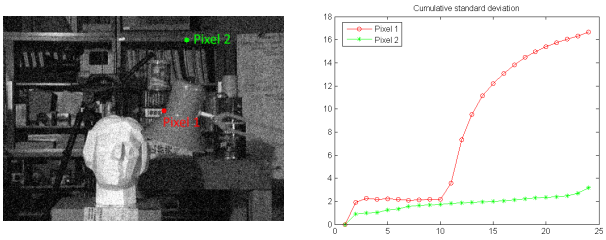


Fig. 3. Cumulative standard deviation (CSD) of pixel 1 (occluded) and pixel 2 (non-occluded). The threshold of CSD for views selection is set to 5. Pixel 1 has 12 views selected for denoising, while pixel 2 has all 25 views selected, including the target view.

Kang et al. [8] introduced the idea of view selection for multi-view stereo where in their situation they were selecting video frames in the temporal dimension, and only implementing a general best 50% selection of all images available. While 50% is a decent percentage for views selection, it lacks flexibility and does not distinguish non-occluded pixels from occluded ones, which eventually downgrades the denoising quality of non-occluded regions as only half of the views are participating in denoising. In this work, we propose to use a dynamic scheme for views selection. An observation of Fig. 2 implies that consistent views should have similar intensity values and image structures, thus leading to a RMSE with lower values and smaller fluctuation. Here we introduce the idea of cumulative standard deviation (CSD), which computes the standard deviation of the sorted RMSE vector cumulatively. Highly inconsistent patches tend to increase the CSD by a significant amount, and by taking only the views with CSD below some threshold, dynamic views selection is achieved. An example of CSD at different locations of the image is shown in Fig. 3, in which we can observe the differences between occluded and non-occluded regions.

2.5. Patch-based Multi-view Non-local Means

The original single view non-local means [2] estimates the target pixel value by weighted averaging pixels within some neighborhood, where the weights depend on the similarities between the matching patches and target patch. Patch-based version of non-local means has also been implemented by Buades et al. [10]. We extend the idea of patch-based non-local means to multiple views by searching similar patches not only in 2D neighborhood, but also across all views. Specifically, we extract patch volumes in the neighborhood of the target patch from the 3D focus image stacks, and apply adaptive window and views selection introduced in section 2.3 and 2.4 to filter out inconsistent patches from each patch volume. The Euclidean distance between the target patch and each patch volume is then computed. With average patch disparities already computed in previous steps, we are selecting patch volumes with closer average patch disparities and smaller Euclidean distances such that a similar patch is both from a similar depth and is of close Euclidean distance to the target patch.

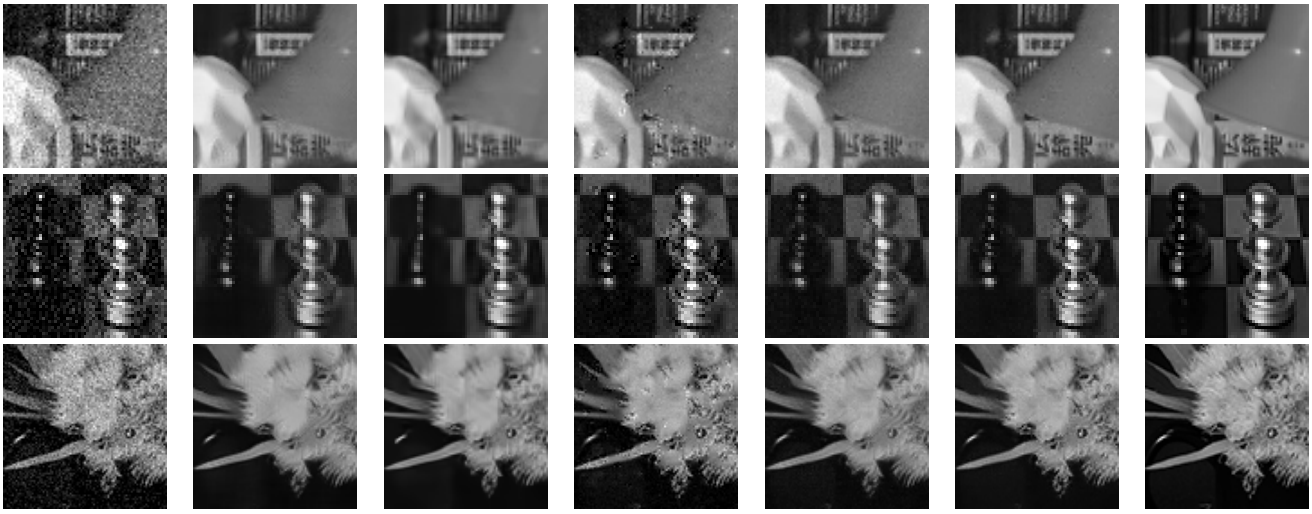


Fig. 4. Comparison of different denoising algorithms. From top to bottom: Ohta, Chess, and Flower image sets; From left to right: noisy image, NLM [2], BM3D [3], Mitaya et al. [6], our previous work [7], proposed method, and ground truth.

With all similar patches in the 3D neighborhood found, we are able to denoise the target patch by weight averaging these patches, where the weight depends on the patch similarities, just as the conventional NLM. A final aggregation step is then taken to form the denoised image from denoised patches. Each pixel is covered by multiple denoised patches, and to determine the value of the pixel in the denoised image, we can take an average of all denoised patches that covers this pixel. This aggregation step makes the algorithm more robust to noise than pixel-wise estimation.

3. EXPERIMENT RESULTS

We implemented our method in MATLAB® R2014a on a machine with Intel® Core™ i7-4700MQ CPU (2.40GHz). Totally six multi-view image sets from Middlebury Stereo Dataset [11] and Stanford Light Field Archive [12] were evaluated. Without loss of generality, we use 5×5 views and set the center view to be the target view. The patch size is set to be 5×5 pixels and a maximum number of 10 similar neighboring patch volumes are selected from a searching area of size 21×21 pixels. For all the image sets, white Gaussian noise of zero mean and variance σ^2 ($\sigma = 20$) were added. Peak signal-to-noise ratio (PSNR) is used as a measuring criteria for denoising quality.

We compared our algorithm with the classical single view denoising methods like non-local means (NLM) [2] and block-matching 3D (BM3D) [3], as well as some of the prior multi-view denoising methods including Miyata's fast multi-view image denoising [6] and our previous work [7]. The denoising results are shown in Table 1, and part of snapshots of the images are displayed in Fig. 4. In general, our methods substantially outperforms the previous methods by a big margin, both visually and quantitatively. This dramatic improvement in denoising quality is attributed to the increased number of patches participating in denoising across multiple views, as well as the exclusion of inconsistent patches achieved by the occlusion handling process.

4. CONCLUSION

We have proposed a new multi-view image denoising algorithm that extends the single-image non-local means to multiple views and is able to handle occlusion elegantly. Quantitative and qualitative experiments with different sets of multi-view images have shown that the proposed method outperforms most of the existing state-of-art algorithms by a great margin, and further proved that the performance limitation of image denoising can be extended by using multiple views. Our future tasks include incorporating the occlusion handling into disparity estimation to further enhance the denoising quality, and improving the computational time of the algorithm.

5. ACKNOWLEDGEMENT

This work was supported by the U.S. National Science Foundation through the Cyber-Physical System program under grant number CNS-1329481.

Dataset	NLM	BM3D	Miyata et al. [6]	Our Previous Work [7]	Proposed
Ohta	29.13	31.41	28.29	31.20	**32.43**
Chess	29.29	30.65	28.45	30.74	**31.87**
Flower	30.20	30.81	29.73	31.01	**32.01**
Knight	27.69	30.20	28.31	30.81	**32.35**
Tarot	25.03	26.38	25.20	27.05	**29.73**
Truck	30.53	32.37	30.57	32.89	**33.88**

Table 1. Denoising quality (PSNR in dB) comparison

6. REFERENCES

[1] L. Zhang, S. Vaddadi, H. Jin, and S. Nayar, "Multiple view image denoising," in *Proc. IEEE CVPR'09*, pp. 1542-1549, Jun. 2009.

[2] A. Buades, B. Coll, and J. Morel, "A non-local algorithm for image denoising," in *Proc. IEEE CVPR'05*, pp. 60-65, Jun. 2005.

[3] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, "Image denoising by sparse 3-D transform-domain collaborative filtering," *IEEE Trans. Image Process.*, vol. 16, no. 8, pp. 2080-2095, 2007.

[4] E. Luo, S. H. Chan, S. Pan, and T. Nguyen, "Adaptive non-local means for multiview image denoising – search for the right patches via a statistical approach," in *Proc. IEEE ICIP'13*, pp. 543-547, Sep. 2013.

[5] Z. Xue, J. Yang, Q. Dai, and N. Zhang, "Multi-view image denoising based on graphical model of surface patch," in *3DTV-Conference*, pp. 1-4, Jun. 2010.

[6] M. Miyata, K. Kadoma, and T. Hamamoto, "Fast multiple-view denoising based on image reconstruction by plane sweeping," in *Proc. IEEE VCIP'14*, pp. 462-465, Dec. 2014.

[7] S. Zhou, Y. H. Hu, and H. Jiang, "Multiple view image denoising using 3D focus image stacks," in *IEEE GlobalSIP'15*, pp. 1052-1056, Dec. 2015.

[8] S. B. Kang, R. Szeliski, and J. Chai, "Handling occlusions in multi-view stereo," in *Proc. IEEE CVPR'01*, vol. 1, pp. I-103, Jun. 2001.

[9] R. T. Collins, "A space-sweep approach to true multi-image matching," in *Proc. IEEE CVPR'96*, pp. 358-363, Jun. 1996.

[10] A. Buades, B. Coll, and J. Morel, "Non-local mean denoising", *Image Process. On Line*, vol. 1, 2001. Available: http://dx.doi.org/10.5201/ipol.2011.bcm_nlm.

[11] Middlebury Multi-view Stereo Datasets. Available: http://vision.middlebury.edu/stereo/data/scenes2001/data/tsukuba/.

[12] The (New) Stanford Light Field Archive. Available: http://lightfield.stanford.edu/lfs.html.