MULTIVIEW 3D VIDEO DENOISING IN SLIDING 3D DCT DOMAIN

[1]*Michal Joachimiak,* [2]*Dmytro Rusanovskyy*

[1]Dept. of Signal Processing,
Tampere University of Technology,
Tampere, Finland

[2]*Miska M. Hannuksela,* [1]*Moncef Gabbouj*

[2]Nokia Research Center,
Nokia Corporation,
Tampere, Finland

ABSTRACT

With the widespread interest in 3D technology areas such as displays, cameras, and processing, the 3D video is becoming widely available. Due to correlation between views in multiview 3D video at the same temporal location, it is possible to perform video processing operations more efficiently comparing to regular 2D video. In order to improve denoising performance for multiview video, we propose an algorithm based on denoising in 3D DCT domain, which is competitive in performance with state-of-art denoising algorithms and it is suitable for real-time implementation. The proposed algorithm searches for corresponding image patches in temporal and inter-view directions, selects 8 patches with lowest dissimilarity measure, and performs denoising in 3D DCT domain. The novel inter-view image patch search method brings up to 1.62dB gain in terms of average luma Peak Signal-to-Noise Ratio (PSNR), with average gain 0.6 - 0.8 dB depending on the amount of noise present in test sequences.

Index Terms— multiview, denoising, 3D, DCT, video

1. INTRODUCTION

Advances in 3D video display and coding, build motivation for capture of high quality 3D video. 3D video comprises at least two views, depicting the same scene from different view points. Displays which are able to reproduce more than two views are already available on the market. At least two types of multiview autostereoscopic displays (ASDs) are available. They can render 9 or 28 views. In practice, that large number of views cannot be transmitted. It is possible to transmit less views and synthesize the missing ones on the receiver side. In order to encode large number of views, the standardization process of the 3D video coding was initiated by the Call for Proposals at the beginning of 2011 [1]. These conditions create a demand for video processing techniques, which can be used alternatively to 2D video processing methods, and which can bring more gain when 3D video is taken into account.

The capture of multiview video is realized using standard video cameras, bundled into the multiview acquisition system. Similarly to 2D video, multiview video acquisition process shares the problem of image noise arising from many sources, mainly, shot noise, dark current noise, fixed pattern noise, amplifier noise, and quantization noise. The presence of noise, not only can degrade perceptual video quality itself, but can significantly affect 2D video processing techniques such as segmentation, object recognition, indexing. In case of 3D video, processing techniques like depth estimation process, based on stereo matching [2], or 3D reconstruction [3] are affected by noise.

There is a broad range of methods for image and video denoising. Recent state-of-art image denoising methods are described in [4] and [5]. Natural extension to image denoising is video denoising, where in addition to spatial correlations, the temporal ones can be taken into account. Several state-of-art approaches were introduced over the last decade including [6, 7, 8]. Among those, [6] is highly computationally optimized and suitable for parallel implementation.

Currently, there exist two approaches to multiple view image denoising. In [9] the problem of imaging with small aperture and short exposure is solved as a denoising problem. Corresponding image patches are found based on a new dissimilarity measure, which takes into account a set of image patches in a reference view, and patches in other views, corresponding to the reference set of patches. Selected patches with similar underlying image structure are denoised using two different methods, Principle Component Analysis (PCA) and tensor analysis. The strengths of this algorithm are using depth map, to get more accurate match, and accurate joint multiview image patch matching. Unfortunately, the computation cost of image patch matching, depth estimation, and methods used for decorrelation is very high, making this method not suitable for real-time application. Different approach, without using depth map, is presented in [10]. As the first step, the images are denoised using BM3D [5] algorithm. After that, the patch-based multi-view stereo (PMVS) [11] model reconstruction algorithm is used to identify feature points, which are projected to other views. These feature points serve as patch centers for image patches, collected for denoising. To calculate the similarity between patches, the graph of surface patches is created. The dissimilarity measure is the geodesic distance calculated between every pair of patches in the graph. The patches with the smallest

dissimilarity measure are selected for Wiener filtering.

Unfortunately, not much attention has been put to 3D video denoising, in case of which, image patches from temporally synchronized views can significantly contribute to denoising. 3D video denoising is different from multi image and 2D video denoising, since it contains both temporal redundancy and inter-view correlations. Thus, it needs different approach than the ones presented before. In this publication, we propose to extend the framework of video denoising in sliding window 3D DCT domain [6] to a multiview case. In addition to temporal direction, an inter-view direction is exploited. Bidirectional block-based motion search selects 8 image patches to create a 3D block volume. On this volume, a sliding window filtering is performed. At single step of the sliding window procedure, a block volume of size 8x8x8 is extracted from the 3D volume, transformed using 3D DCT transform, hard-thresholded, and inverse transformed. As far as the authors are aware, the proposed solution is the first approach to 3D video denoising. It is computationally relatively lighweight, uses DCT transform, which is widely spread in modern image and video coding standards, and does not require depth map estimation, which is computationally costly.

The paper is organized as follows. Section 2 describes video denoising using Sliding Window 3D DCT algorithm (SW-3DDCT). In Section 3 a multiview extension to SW-3DDCT is proposed. Section 4 depicts experimental results. Section 5 contains relevant conclusions and discussion.

2. DENOISING IN SLIDING WINDOW 3D DCT DOMAIN

The goal of the Sliding Window 3D DCT (SW-3DDCT) video denoising process is to restore a set of original frames $\mathcal{I}(t_r) = \{I(t)\}_{t=1}^{N}$, from the set of noisy observations $\mathcal{Y}(t_r) = \{Y(t)\}_{t=1}^{N}$. The 3D volume is created from the set of 2D image patches, selected from the set of N frames at consecutive temporal locations, centered at the reference temporal location t_r. Within one 3D volume the temporal locations are selected following the equation:

$$t_\mathrm{r} - N/2 \leq t < t_\mathrm{r} + N/2 \tag{1}$$

The reference frame is traversed along vertical and horizontal directions. For each location in the reference frame a separate 3D volume is created. To achieve high correlation between image patches, an image patch matching process - motion estimation is used. Effectively, the 3D volume size is $W \times H \times N$, where W and H are width and height of the image patch, and N is the number of patches from temporal positions. The process of 3D volume creation is depicted in Fig. 1.

High correlation between patches in the 3D volume is exploited by 3D DCT transform [6]. In vertical and horizontal directions the 2D transform is performed, whereas in temporal direction 1D DCT is applied. The 3D volume block size is larger than the DCT transform size. As a consequence the 3D-DCT transform operates on the 3D volume in a sliding manner. Every denoising step on the 3D volume is composed of the following consecutive parts:

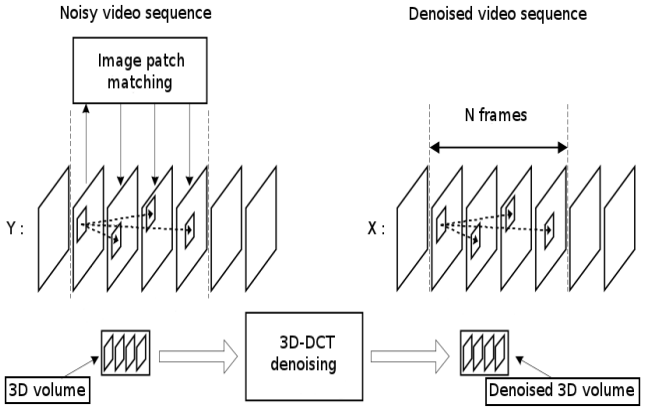

Fig. 1. The 3D volume creation and SW-3DDCT denoising process. In the step depicted, frames between vertical dashed lines participate in denoising process.

- 3D-DCT forward transform,
- Hard thresholding of 3D-DCT coefficients with weight accumulation in temporal direction for nonzero coefficients,
- Inverse 3D-DCT transform,
- Storing weighted inverse transformed values to all image patches in the volume.

Denoising is executed by shrinkage in the 3D transform domain. The procedure of hard thresholding is used to eliminate transform coefficients, corresponding to high frequency components, containing the noise.

3. MULTIVIEW SW-3DDCT

In case of multiview video denoising, the aim is to restore original set of frames

$$\mathcal{I}(t_r, m_r) = \{I(t, m_r)\}_{t=1}^{N} \cup \{I(t_r, m)\}_{\mathrm{m=1}}^{M} \tag{2}$$

from its corresponding, noisy observations. The idea to extend the SW-3DDCT filtering was based on its competitiveness to the other state-of-art denoising methods [12] and, the fact that, vast range of video processing tools uses similar functionalities, already implemented in software and hardware, namely, motion estimation and DCT transform. The process described below refers to luma component only. For chroma components the procedure is repeated.

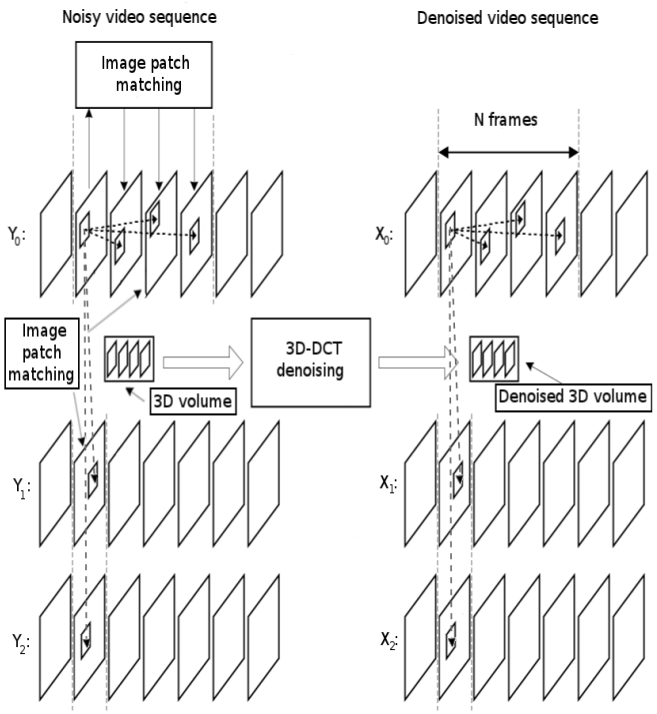

Fig. 2. The 3D volume creation for multiview denoising. Image patches are searched in frames between dashed vertical lines.

The set of frames is compound of temporal window of size N, centered at the reference temporal location t_r of the reference view m_r. In addition, frames from all other $M-1$ views, at the reference temporal location t_r are taken into account, during image patch matching. To decrease the complexity, the fast, coarse-to-fine motion estimation is used. It was found, that full search motion estimation performance was approximately equal to course-to-fine. At the first step, candidates at locations sampled every 8 pixels are checked. After the best candidate is found, the search window sampling is decreased to 4. The procedure is repeated until integer pixel accuracy is reached.

In our algorithm, the 3D volume is created out of 8 image patches of size 16×16, from temporal and inter-view directions. The dissimilarity measures for all patches are stored and later sorted, according to the smallest values. Only 8 image patches, with highest similarity, are selected for denoising. The heap sort algorithm is used for sorting. To improve the sorting speed, it is stopped early, before all the values are sorted, when the number of patches selected for 3D volume reaches 8. On the 3D volume prepared in that way, the regular SW-3DDCT procedure is employed. The process of multiview SW-3DDCT denoising is exposed in Fig. 2.

4. EXPERIMENTAL RESULTS AND DISCUSSION

The algorithm presented in Section 3 was tested on a set of 3D video sequences. Video sequences selected for the experiment were "Ghost Town Fly", "Undo Dancer", "Ghost Town" and "Sneakers" [13]. The first three sequences are computer-generated synthetic videos and they do not contain noise related to an image acquisition process. The "Sneakers" sequence contains computer-generated graphics on a natural image background. Every test sequence contains 9 views, taken at 9 parallel viewpoints. Viewpoints are located on the baseline in an uniform manner.

For the purpose of experiment, the test sequences are corrupted by additive white Gaussian noise (AWGN) $\eta(t) = N(0, \sigma^2)$, with zero mean and variance σ^2. For each temporal location we have:

$$Y(t) = I(t) + \eta(t) \tag{3}$$

For every multiview sequence, four new sequences were created, adding noise with four different standard deviations: 5, 10, 15, 20. Since the standard deviation of the noise can be estimated from the signal [14], hard thresholds for shrinkage operator are assumed to be known a-priori.

Currently available ASDs can render up to 28 views from a subset of views transmitted. The MPEG suggested to use 3-view case [1] to meet the needs of many today's ASDs. In general the number of views in 3D video is not limited. To test the benefit of using additional views, in the case of multiview SW-3DDCT, the experiments were performed on the same noisy sequences for 3 different cases. For each case, the amount of views participating in denoising was different, particularly 3, 5, and 9. The selection was done with the assumption, that the viewpoints location on the baseline should be uniform, regardless the amount of views participating in denoising. For 3-view case, the most distant views were used. For 5-view case, the views from 3-view case were used and views between them, located on equidistant baseline positions. For 9-view case, all available views were used.

For comparison, the performance of the single view SW-3DDCT algorithm [6] was tested. In this case, each view from all sequences in the test set was denoised separately, using SW-3DDCT. The size and amount of patches were the same as in multiview case. Denoising performance was evaluated in terms of luma PSNR, computed against the original, noise-free test sequences. For visual comparison, some fragments of frames extracted from test sequences are depicted in Fig.3. The numerical results of denoising are presented in Table 1. The values present an average PSNR gain over all denoised views. As seen from Table 1, joint denoising of multiple views provides significant improvements compared to a single view denosing. An increase in number of views participating in denoising leads to a performance increase. It can be noticed from the Table 1, that the multiview SW-3DDCT improvement over a single view SW-3DDCT is maximal for

Table 1. Denoising performance of the Multiview SW-3DDCT for 3 different number of views participating in denoising. The performance of the single view SW-3DDCT shown as a reference. The results are presented in terms of average luma PSNR measure over all views participating in denoising.

			PSNR gain				ΔPSNR gain over single view		
Sequence	σ	noise	single view	3 views	5 views	9 views	3 views	5 views	9 views
Ghost Town Fly	5	34.14	38.01	38.27	38.65	39.29	0.25	0.64	1.27
	10	28.13	35.89	36.23	36.74	37.46	0.34	0.85	1.57
	15	24.62	34.21	34.58	35.13	35.83	0.37	0.92	1.62
	20	22.14	33.29	33.65	34.16	34.77	0.36	0.87	1.48
Undo Dancer	5	34.14	37.18	37.19	37.34	37.73	0.01	0.16	0.55
	10	28.15	34.78	34.81	35.03	35.49	0.03	0.25	0.71
	15	24.72	33.06	33.09	33.34	33.83	0.03	0.29	0.77
	20	22.33	32.08	32.10	32.36	32.82	0.02	0.28	0.74
Ghost Town	5	34.14	39.72	39.85	40.05	40.31	0.13	0.33	0.59
	10	28.14	37.39	37.57	37.78	38.05	0.18	0.39	0.66
	15	24.65	35.45	35.63	35.82	36.07	0.18	0.37	0.62
	20	22.21	34.26	34.42	34.58	34.79	0.16	0.33	0.53
Sneakers	5	34.14	42.27	42.28	42.39	42.49	0.01	0.12	0.22
	10	28.13	39.89	39.87	40.05	40.23	-0.01	0.16	0.34
	15	24.61	37.59	37.53	37.70	37.89	-0.06	0.11	0.30
	20	22.12	35.95	35.86	36.00	36.17	-0.09	0.05	0.22
AVERAGE	5	34.14	39.30	39.40	39.61	39.96	0.10	0.31	0.66
	10	28.14	36.99	37.12	37.40	37.81	0.13	0.41	0.82
	15	24.65	35.08	35.21	35.50	35.90	0.13	0.42	0.83
	20	22.20	33.89	34.01	34.27	34.64	0.11	0.38	0.74

"Ghost Thown Fly" sequence, where it brings up to 1.62 dB. The "Ghost Town Fly" sequence contains camera zooming, which breaks block-based temporal correlation. Thus, multiview SW-3DDCT, which exploits inter-view correlations was much more effective in this case. Denoising performance increased together with the number of views participating in the denoising. The average use of interview patches, shown in the Table 2, grows significantly when the number of total views increases. From both tables, it can be seen that increase in use of interview image patches had positive effect on denoising performance increase.

The changes at temporal direction in standard video sequence are mainly due to inter-frame motion or camera movement. This holds for the temporal changes in multiview videos. However, differences between viewpoints are due to different positions of cameras, observing the same scene. With known camera geometry and depth information, improved image patch matching, using warping, can be performed which, most likely, would improve the denoising performance. However, the depth information is not always available and in case the depth is generated from noisy video data, its quality is decreased. Low quality depth map imposes low quality image patch matching, when warping is applied. The advantage of the approach presented here is, that no depth information is needed to perform denoising. In addition, depth map estimation and Depth Image Based Rendering, used for warping image patches, are computationally costly and prevent real-time operation.

Table 2. Average use of interview image patches as a percentage of all patches.

σ	3 views	5 views	9 views
5	9.42%	19.18%	30.95%
10	10.27%	19.82%	31.37%
15	11.02%	20.48%	31.97%
20	11.73%	21.21%	32.72%

5. CONCLUSIONS

We proposed a multiview video denoising method. Its performance is comparable with state-of-art algorithms. Numerical complexity is much lower than [8] since only one patch matching step and no Wiener filtering are used. The DCT transform is widely used and hardware implementations exist. Thus, our method is suitable for real-time implementation. It was tested on multiview 3D video sequences, corrupted by AWGN. Experimental results showed substantial improvement over the existing video denoising method [6]. Gains up to 1.62 dB in terms of average luma PSNR were observed, with average gain from 0.6 to 0.8 dB, depending on the magnitude of noise used to corrupt test sequences.

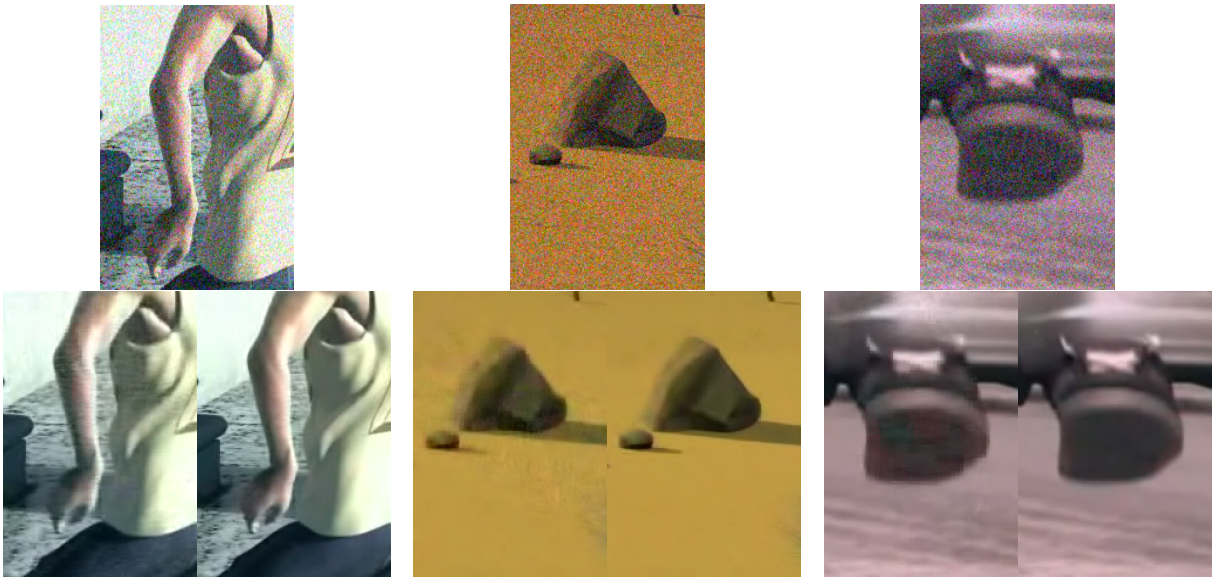

Fig. 3. The results of denoising. Top row contains noisy data with $\sigma = 15$. Bottom row contains pairs of corresponding denoised data, with single view denoising on the left and multiview denoising on the right, within each pair.

6. REFERENCES

[1] ISO/IEC JTC1/SC29/WG11, "Applications and requirements on 3d video coding, doc. n12035," Geneva, Switzerland, March 2011.

[2] P. Leclercq and J. Morris, "Robustness to noise of stereo matching," in *Proc. of the 12th Int. Conf. on Image Analysis and Processing*, Sept. 2003, pp. 606–611.

[3] Z. Xue, J. Yang, Q. Dai, and N. Zhang, "Multi-view image denoising based on graphical model of surface patch," in *3DTV-Conference: The True Vision - Capture, Transmission and Display of 3D Video (3DTV-CON), 2010*, june 2010, pp. 1 –4.

[4] A. Baudes, B. Coll, and J.M. Morel, "A review of image denoising algorithms, with a new one," *SIAM Journal on Multiscale Modelling and Simulation*, vol. 4, pp. 490–530, 2005.

[5] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, "Image denoising by sparse 3-d transform-domain collaborative filtering," *IEEE Trans. on Image Processing*, vol. 16, pp. 2080 –2095, 2007.

[6] D. Rusanovskyy and K. Egiazarian, "Video denoising algorithm in sliding 3d dct domain," in *Proc. of the 7th Int. Conf. on Advanced Concepts for Intelligent Vision Systems (ACIVS)*, 2005, pp. 618–625.

[7] A. Buades, B. Coll, and J-M. Morel, "Nonlocal image and movie denoising," *Int. Journal of Computer Vision*, vol. 76, pp. 123–139, 2008.

[8] K. Dabov, A. Foi, and K. Egiazarian, "Video denoising by sparse 3d transform-domain collaborative filtering," in *Proc. 15th European Signal Processing Conference, EUSIPCO*, 2007.

[9] Li Z., S. Vaddadi, H. Jin, and S.K. Nayar, "Multiple view image denoising," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, CVPR*, 2009, pp. 1542 –1549.

[10] Z. Xue, J. Yang, Q. Dai, and N. Zhang, "Multi-view image denoising based on graphical model of surface patch," in *Proc. of the 3DTV-Conference: The True Vision - Capture, Transmission and Display of 3D Video*, 2010, pp. 1 –4.

[11] Y. Furukawa and J. Ponce, "Accurate, dense, and robust multi-view stereopsis," in *Proc. the Conf. on Computer Vision and Pattern Recognition*, 2007, pp. 1 –8.

[12] G. Varghese and Zhou Wang, "Video denoising based on a spatiotemporal gaussian scale mixture model," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 20, no. 7, pp. 1032–1040, july 2010.

[13] D. Rusanovskyy M. Hannuksela, *"Extension of existing 3DV test set toward synthetic 3D video content"*, ISO/IEC JTC1/SC29/WG11 MPEG2011/ M19221, Daegu, Korea, Jan. 2011.

[14] D. Donoho and I. M. Johnstone, "Adapting to unknown smoothness via wavelet shrinkage," *Journal of the American Statistical Association*, vol. 90, pp. 1200–1224, 1995.