3D Wavelet encoder for depth map data compression

Miguel Martínez-Rach, Otoniel López-Granado, Pablo Piñol, Manuel P. Malumbres
Miguel Hernández University
Avda. Universidad s/n
Elche, Alicante, 03202, Spain
{mmrach,otoniel,pablop,mels}@umh.es

Abstract: Depth-Image-Base Rendering is an effective approach for 3D-TV as it uses the current TV infrastructure, so a 3D video sequence can be rendered at the user-end device with significant bandwidth savings, since it avoids the need of transmitting two video channels, the left and right views. However, quality and time consistence of the depth map sequences is a problem in this field. We present an intermediate solution, for compressing the depth information, that is set in a middle point between using pure INTRA or INTER encoders that it is able to cope with the quality and time consistency of the captured depth map info. Our 3DLTW-D encoder uses the 3D-DWT transform to efficiently encode static depth information in the scene, with reduced encoding/decoding delays and achieving the same visual quality than the current H264/AVC and x264 encoders running in INTRA mode.

1. Introduction

The Depth-Image-Based Rendering (DIBR) [1] is as an effective approach for free-viewpoint three-dimensional television (3D-TV) whose main advantage, compared with traditional representation of 3D video, based on left and right views, is that provides high quality virtual views with lower bandwidth requirements, because the depth map data can be more efficiently encoded that natural images. So, instead of encoding two separated image streams (left and right views), with the DIBR approach only one view and the corresponding depth map will be encoded and decoded.

A major problem with the DIBR method is that high quality and time consistent depth maps are required for proper reconstruction. In order to capture depth information, low-resolution range sensors based on time-of-flight (ToF) principles or laser scanners are used. These methods [2] produce precise depth measurements but rather noisy and with lower resolution than the corresponding color image.

Depth maps can be obtained too, using "depth-from-stereo" or "depth-from-multiview" algorithms, based on finding disparities between images from nearby cameras and converting them to the corresponding depths [3]. This kind of algorithms suffer from inaccuracies in finding the disparity correspondences and introduce temporal variations over the depth values belonging to pixels from static objects or inclusive from the background of the scene. The quality of the delivered depths also varies from approach to approach.

In addition, the depth map data compression introduces additional errors like blocking artifacts when DCT (Discrete Cosine Transform) based encoders are used [4], or ringing artifacts when DWT (Discrete Wavelet Transform) based encoders are used.

It is important then, to process and compress depth maps in a way that it minimizes distortion in the final rendered virtual views. Enhancing the quality of the original depth map jointly with an efficient compression strategy, a post processing stage for reducing

compression artifacts (i.e. blocking, blurring, ringing, etc), and the use of an optimized DIBR algorithm are the key points for getting high quality rendered virtual views.

Recently, several depth map compression methods have been proposed. These methods can be broadly classified in two categories. The first one proposes radical compression techniques such as platelet based coding [5], silhouette based coding [6] and 3D motion estimation based methods [7]. The other one tries to optimize existing video codecs to properly encode depth maps, such as novel methods for mode selection [8] and reconstruction filters [9].

Several filtering techniques [10,11,12,13] are proposed in order to reduce the impact of spatial and temporal inconsistences in the depth maps that finally results in artifacts in the rendered views. In [9,11,14,15] bilateral filtering or joint bilateral filtering is used for edge-preserving the important structural information used in the DIBR algorithm, and at the same time performing a smoothing of the depth values resulting in higher quality rendered views.

Temporal inconsistences between successive frames in a video sequence are based on smooth depth variations in the depth values for pixels belonging to objects located at static areas of the scene or just in the background. These variations produce a flickering effect in the rendered sequence, in areas that the viewer expected to be static, modifying therefore the subjective perception of the overall sequence quality. Temporal smoothing or median filtering between successive frames [16,17,18] can be applied to these areas in order to reduce this flickering effect in the final reconstructed video sequence.

Optimal depth map video sequences should exhibit a strong temporal consistence in the depth values for the scene background and for objects that are static, i.e. at the same distance from the camera for a given time interval.

The aim of this paper is based in this idea, so we introduce the use of a 3D-Wavelet transform based encoder, with a preprocessing stage that reduce the temporal inconsistence of the original depth map and a post processing stage that reduces the ringing artifacts introduced by the wavelet transform at high compression rates, in the same way that DCT encoders have a post processing de-blocking and de-blurring stage.

The goal is to achieve the best R/D performance at high compression rates of the depth map by exploiting the ability of the 3D transform in compacting temporal information in low temporal frequency subbands. The enhanced, compressed and filtered depth maps and the undistorted color view are served as input to the DIBR process [19], so the impact of the depth map compression strategy can be properly measured in the final reconstructed views.

The use of a 3D-Wavelet transform set our proposal in an interesting intermediate point between INTER frame encoders, that uses motion estimation and motion compensation requiring high computational cost, and pure INTRA frame encoders like JPEG2000 used in a frame-by-frame basis. Results show that this proposal will be well suited for high frame rate video sequences (30 fps or more) or low and moderately low motion scenes.

Quality performance of rendered views and computational cost for our proposed **3DLTW-D** (**3D L**ower **T**ree **W**avelet for **D**epth-maps) encoder are compared with H264/AVC and x264, both in INTRA mode, with the highest quality mode configuration, i.e. exploiting all the spatial prediction modes available. Also, the native LTW INTRA [20] frame encoder will be used to measure the real impact of the temporal transform.

2. Depthmap processing

In order to improve the temporal consistency among depth map, we use a simple but fast temporal smoothing algorithm over the depth values. The idea behind this pre-processing step is to assert that pixels belonging to static objects should have exactly the same depth value along time, since their distance to the camera does not change. For each pixel belonging to static objects or background, the median of its depth values along time is calculated. The median value is assigned to its corresponding pixel, removing the small depth variations found along the temporal dimension in a specific time interval.

Some segmentation problems may arise when facing this task. Static objects and its corresponding pixels should be detected in the scene. A moving object invalidates the depth values of static objects and background pixels when its movement occludes them. In addition, if the scene contains large areas of moving objects or motion affects to the whole scene, the amount of static "clean" pixels becomes smaller respect those "dirty" pixels whose depth has change some time in the scene due to movement.

A proper temporal window size should be fixed, i.e. the amount of frames, for which the pixels belonging to static objects and background remain unaltered and could be smoothed. Our experiments produce better results with a small temporal window than with a larger one, because the amount of static pixels is higher. In [18] authors analyzed the most appropriate temporal window size for depth map processing, concluding that a temporal window size of 5 frames will be the best option for these tasks. So, we will also employ a temporal window size of 5 frames. This window size also corresponds to the size of the temporal 3D-DWT window fact that improves the performance of the incoming 3D-Wavelet temporal filtering.

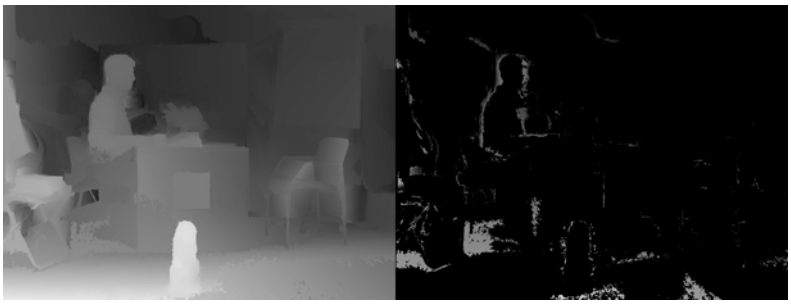


Figure 1: Smoothing area of first temporal window of the Book Arrival Sequence.

Segmentation or motion estimation approaches could be applied to detect the pixels belonging to static objects and background in a window, but they are time consuming tasks and we would like to proceed at the same time that the 3D-Wavelet transform is applied.

We calculate the Mean Absolute Deviation (MAD) of depth value of each pixel for the actual temporal window. Those pixels whose MAD value is lower than a threshold are classified as "clean" pixels and their values are set to the median value found in the temporal window.

The left image of Figure 1 shows a frame of the first temporal window from the Book Arrival sequence. The right image shows a frame composition where black pixels correspond to pixels having MAD value lower than that threshold. For that frame the

96.33% of pixels are labeled as “clean” pixels and therefore their depth value is set to the median value of all pixel depths in the temporal window.

Figure 2: Ringing effect and result after the application of two dimensional joint bilateral filter to the 2D wavelet compressed depth map at 0.1 bpp

As the quality of the depth map and the edge-structural information is directly correlated with the final quality of the rendered views after the DIBR algorithm, it is very important to remove this ringing effect and at the same time to preserve the border information. For that purpose, we use the joint bilateral filter presented in [9] as deringing filter for the post-processing stage. In Figure 2 we see the ringing effect that the 2D wavelet transform introduces in the depth map. Left image corresponds to the original depth map, central image corresponds to the reconstructed depth map at 0.1 bpp where ringing artifacts are clearly present, and right image corresponds to the reconstructed and post-processed depth map with the selected filtering strategy that removes this ringing effect.

3. 3DLTW-D

3DLTW-D encoder is based on a frame-by-frame 3D-DWT scheme [21] and lower trees with eight descendants.

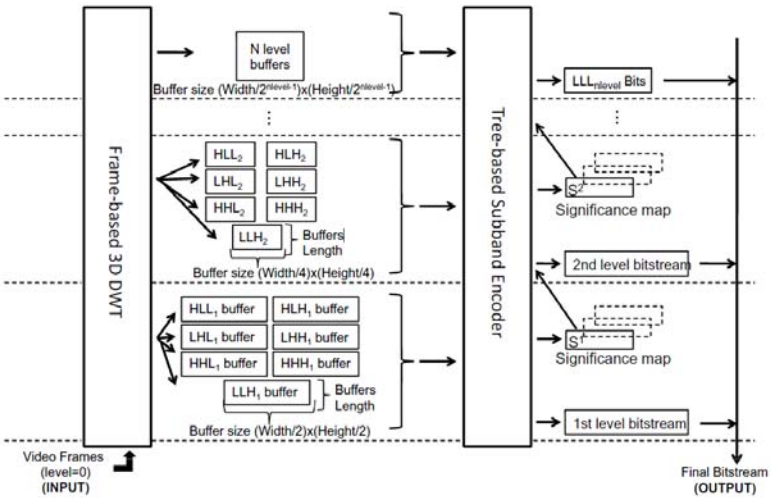


Figure 3: Overview of the proposed tree-based encoder with efficient use of memory

Figure 3 shows the overall system. The 3D-DWT module releases subband frames at different decomposition levels. At each level the subband frames are stored in a dedicated encoder buffer. There are two subband frames for each subband type. When this buffer is full, the 3D-DWT encoder processes all subbands and maintains the significance map for building the trees.

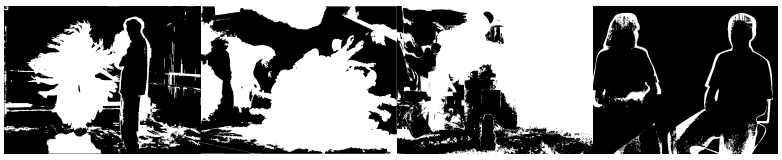
The 3DLTW-D encoder has to determine if each 2x2 block of coefficients of both subband frames stored in the encoding buffer is part of a lower-tree. If the eight coefficients in these blocks are lower than the quantization threshold, and their descendant offspring are also insignificant, they are part of a lower-tree and do not need to be encoded. In order to know if their offspring are significant, we need to hold a binary significance map of every encoder buffer (SL in the figure) because the encoder buffer is overwritten by the wavelet transform once it is encoded, and hence the significance for their ascendant coefficients is not automatically held. The significance of both 2x2 blocks can be held with a single bit.

When there is a significant coefficient in both 2x2 block or in its descendant coefficients, we need to encode each coefficient separately. Recall that in this case, if a coefficient and all its descendants are insignificant, we use the LOWER symbol to encode the entire tree, but if it is insignificant, and the significance map of its eight direct descendant coefficients shows that it has a significant descendant, the coefficient is encoded as ISOLATED_LOWER. Finally, when a coefficient is significant, it is encoded with a numeric symbol along with its significant bits and sign.

4. Results

Simulations are performed for the test sequences presented in Table 1. Analyzing the characteristics of the different depth sequences used in the test, we classify them in terms of frame rate and motion activity, taking into account not only the velocity of objects in the scene but also the amount of the frame area covered by motion along the sequence.

These two parameters have a huge impact on the capacity of the 3D-Wavelet transform to compact temporal frequencies. The smoother the depth variation between frames is, the better the 3D-Wavelet transform behaves. Of course the quality of the depth map sequence is also a key point in this behavior. As said before, in optimal depth map sequences, static objects will have the same depth value in every frame.



a) Ballet b) Breakdancers c) Book Arrival d) Interview

Figure 4: Static (black) and motion (white) areas of depth map sequences

Table 1: Classification of depth sequences by type of motion

Sequence	Frame Size	Motion degree	Frame rate (fps)	Frame occlusion
Ballet	1024x768	Fast	15	30.94 %
Breakdancers	1024x768	Very Fast	15	57.46 %
Book Arrival	1024x768	Fast	30	62.39 %
Interview	640x512	Very Slow	30	9.21 %

In Figure 4 we can see how motion affects to different frame areas for each sequence, leaving more pixels unaffected by motion and therefore classified as background pixels (black) whose temporal depth information will be better compacted by the 3D-Wavelet transform in temporal low frequency coefficients. These background pixels include the

pixels belonging to objects whose depth does not vary (over a threshold) in the whole sequence.

So, the best sequence to be compressed with the temporal transform is “Interview” because the motion of objects is very slow and the frame rate is high resulting in small occluded frame areas. The worst sequence is therefore “Breakdancers”, because frame occlusion size is high, the motion of objects in the scene is very fast and the frame rate is low. The other two sequences are just in the middle point with different frame rates and frame occlusion areas.

For each original test sequence, the depth map was encoded and decoded with the 3DLTW-D. The encoder smoothed temporally the input depth sequence in the way exposed previously. Then, de-ringing step by means of the joint bilateral filter proposed in [9] was applied to the reconstructed sequence in the decoder. Resulting depth sequences are the input to the DIBR [19] algorithm jointly with the original color sequences.

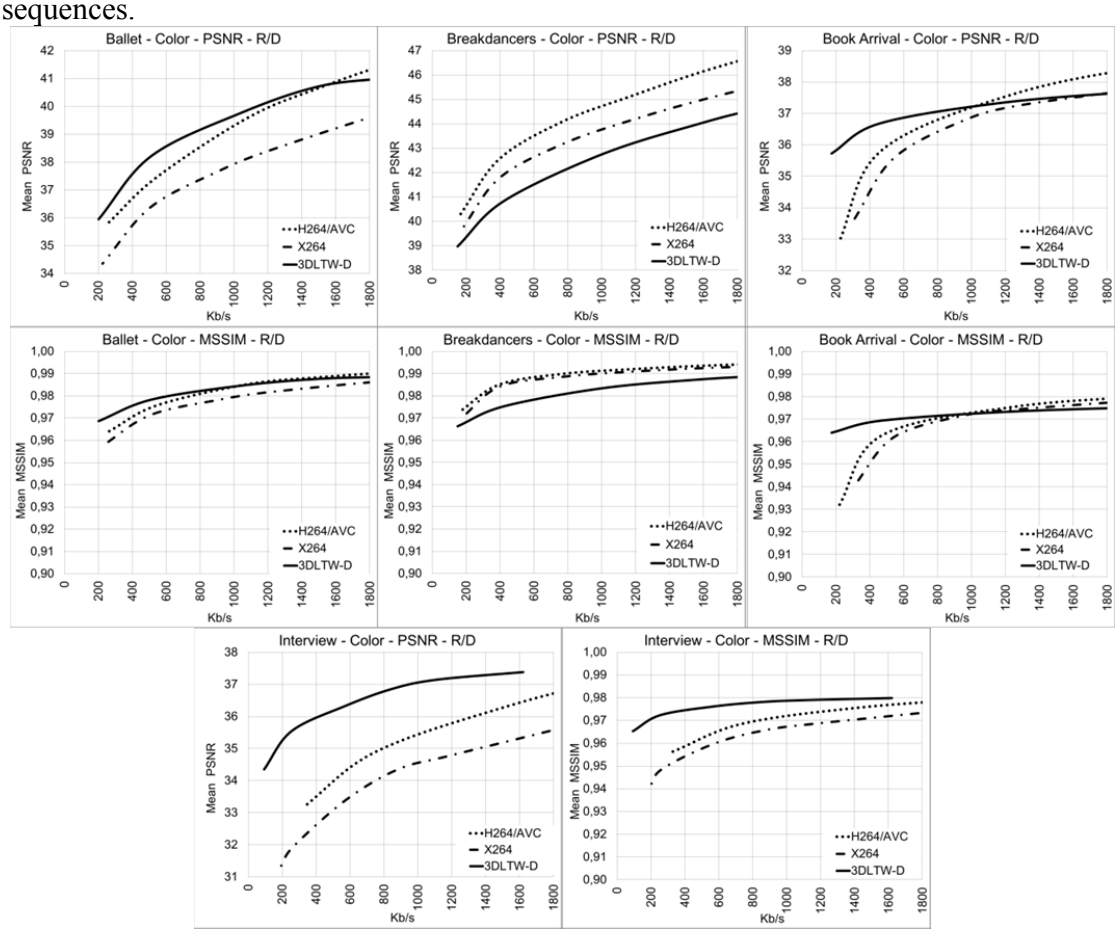


Figure 5: R/D for the test sequences. Quality is shown in terms of PSNR, and in terms of the MSSIM quality metric as labeled in each graph.

The results are compared to x.264/Intra (FFmpeg version SVN-r25117, High profile, level 4.0) and H.264/AVC/Intra (High-10, JM16.1) with all intra-prediction MB modes enabled. Our encoder does not include any intra or inter prediction mode. As described

before only the temporal wavelet transform is applied as an intermediate position between inter and pure intra encoders.

The rate/distortion behavior of the rendered views is compared in terms of PSNR and the MSSIM [22]. Although most studies employ PSNR metric to measure video quality performance, we decided to use in our study an objective quality assessment metrics too. There are several studies about the convenience of using other video quality metrics than PSNR in order to better fit to human perceptual quality assessment (i.e subjective tests) [23,24,25,26].

Figure 5 shows the rate distortion curves where quality is measured in terms of PSNR and in terms or MSSIM. They represent the mean quality value for the left and right rendered views. The vertical axis shows the quality and the horizontal axis the compression rate which corresponds to the depth map sequence. The original color sequence is not compressed so we can measure the effect of the compression of the depth map in the final quality of the rendered views.

a) H264/AVC b) x264 c) 3DLTW-D

Figure 6: Cropped left rendered view for: Breakdancers at 760Kb/s and Interview at 300 Kb/s.

PSNR values do not saturate as the rate increases and in contrast, this saturation occurs when measuring quality with the MSSIM metric. This effect is well known [24] and in our experiments the perceived quality of the rendered views does not increase above certain rate that depends on the sequence.

For Breakdance the performance in terms of PSNR of our 3DLTW-D encoder is far from x264 and H264/AVC. But perceptually, in terms of MSSIM, that difference is not as big. Although the difference in PSNR between x264 and H264/AVC is up to 0.95 dBs at 760 Kb/s, this difference is practically inappreciable at this rate, and in the rest of the curve, when the MSSIM metric is used. At the same rate, difference between 3DLTW-D and H264/AVC is 1.9 dBs (PSNR) but only 0.0095 MSSIM points. First row of Figure 6 shows the Breakdance sequence for each codec where no subjective difference is noticeable at this rate.

However, for the Interview sequence differences in PSNR between 3DLTW-D, H263/AVC and x264 at 300 Kb/s are 2.5 dBs and 3.3 dBs respectively for the left

rendered view corresponding to 0.018 points and 0.023 MSSIM points respectively. These differences are noticeable as shown in second row of Figure 6. The maximum differences between the three codecs are resumed in Table 2, where positive values represent a gain of quality when 3DLTW-D is used, and negative values corresponds to a loss of quality. These values correspond to the low end of the rate range where an extreme compression of the depth map is achieved. At higher rates, from 400 Kb/s to 1200 Kb/s the perceptual quality of the rendered views is almost the same, according to the MSSIM metric values.

Table 2: Maximum PSNR and MSSIM differences between encoders. Negative values corresponds to a loss of quality

3DLTW-D vs.	**H264/AVC**		**X264**	
	PSNR (dB)	MSSIM	PSNR (dB)	MSSIM
Ballet	1.08	0.006	2.10	0.012
Breakdancers	- 1.95	- 0.009	- 1.03	- 0.008
Book Arrival	3.10	0.034	2.70	0.027
Interview	2.60	0.016	3.60	0.024

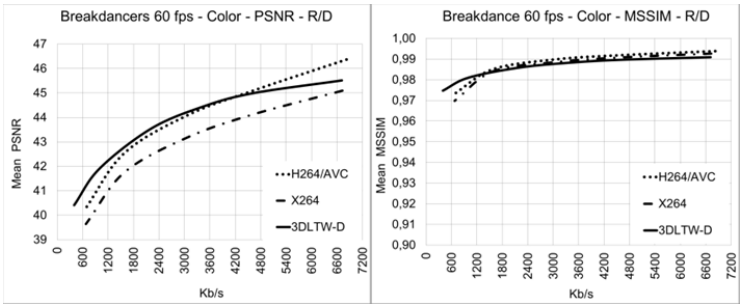


Figure 7: R/D for the Breakdancers sequences at 60 fps.

For the Breakdance sequence at 15 fps, the 3DLTW-D encoder obtains the worst results because differences between consecutive frames are high due to fast motion in the scene. But if we increase the frame rate up to 60 fps, the 3DLTW-D encoder achieves the same quality as the H264/AVC regardless the quality metric being used (see Figure 7).

Table 3 records the encoding times for the three evaluated encoders. Only the temporal smoothing pre-process step is included as the rest of the process, i.e. the spatial filtering and the DIBR process is done at the decoder side.

Table 3: Execution time comparison of the encoding process

Frame size	**H264/AVC**	**X264**	**3DLTW-D**
1024x768	0.54 fps	18 fps	27.94 fps
640x512	1.11 fps	24 fps	77.90 fps

5. Conclusions

The 3DLTW-D encoder presented in this work uses the 3D-Wavelet transform to compress the depth maps of multiview or 2D + depth sequences as an intermediate solution between using pure INTRA and INTER encoders.

Our 3D-Wavelet encoder is well suited for high frame rate sequences obtaining the same visual quality than the current standards running in INTRA mode, and being much faster, up to 9 fps in 1024x768 frame size and 53 fps for 650x512 frame size.

As presented in this work, the proper use of temporal smoothing strategies of the depth sequences, a post-processing deringing filter and the encoding time and quality performance that exhibits our 3D-Wavelet encoder in comparison with the H264/AVC and x264 encoders sets this solution as an attractive one for high frame rate sequences.

References

[1] Christoph Fehn, "Depth-Image-Based Rendering (DIBR), Compression and Transmission for a New Approach on 3D-TV," Proc.of the SPIE, vol. 93 & 5291, 2004.

[2] J. Zhu, L. Wang, R. Yang and J. Davis, "Fusion of Time-of-Flight Depth and Stereo for High Accuracy Depth Maps", IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), 2008.

[3] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms.," International Journal of Computer Vision, vol. 47, pp. 7-42, April-June 2002.

[4] P. Merkle, Y. Morvan, A. Smolic, D. Farin, K. Mueller, P. H. N. de With, T. Wiegand, "The effects of multiview depth video compression on multiview rendering", Signal Processing: Image Communication, January 2009.

[5] Y. Morvan, D. Farin, and P. H. N. de With, "Depth-Image Compression Based on an R-D Optimized Quadtree Decomposition for the Transmission of Multiview Images," in IEEE International Conference on Image Processing, 2007, vol. 5, p. V - 105-V - 108 ICIP 2007.

[6] S. Milani and G. Calvagno, "A Depth Image Coder Based on Progressive Silhouettes," Signal Proc. Letters, IEEE, vol. 17, no. 8, pp. 711-714, 2010.

[7] D. V. S. X. De Silva, W. A. C. Fernando, and S. L. P. Yasakethu, "Object based coding of the depth maps for 3D video coding," Consumer Electronics, IEEE Transactions on, vol. 55, no. 3, pp. 1699-1706, 2009

[8] D. V. S. X. De Silva and W. A. C. Fernando, "Intra mode selection for depth map coding to minimize rendering distortions in 3D video," Consumer Electronics, IEEE Transactions on, vol. 55, no. 4, pp. 2385-2393, 2009.

[9] De Silva, D.V.S.X.; Fernando, W.A.C.; Kodikaraarachchi, H.; Worrall, S.T.; Kondoz, A.M.; , "Improved depth map filtering for 3D-TV systems," Consumer Electronics (ICCE), 2011 IEEE International Conference on, pp.645-646, 9-12 Jan. 2011

[10] Y. Qingxiong, Y. Ruigang, J. Davis, and D. Nister, "Spatial-Depth Super Resolution for Range Images," in CVPR'07, 2007.

[11] J. Kopf, M. Cohen, D. Lischiski, M. Uyttendaele, "Joint Bilateral Upsampling', in Proc. SIGGRAPH conf., ACM Trans. Graphics, 26(3), 2007.

[12] A. K. Riemens; O. P. Gangwal; B. Barenbrug; R.-P. M. Berretty, _Multistep joint bilateral depth upsampling', in Proceedings of SPIE, Vol. 7257, Visual Communications and Image Processing 2009, Majid Rabbani; Robert L. Stevenson, Editors, pp. 72570M.

[13] S. Smirnov, A. Gotchev, and K. Egiazarian, "Method for Restorations of Compressed Depth Maps: A Comparative Study," in VPQM2009, 2009.

[14] C. Tomasi and R. Manduchi, "Bilateral Filtering for Gray and Color Images," in IEEE International Conference on Computer Vision, Bombay, 1998.

[15] Gangwal, O.P.; Berretty, R.-P.; "Depth map post-processing for 3D-TV," Consumer Electronics, 2009. ICCE '09. Digest of Technical Papers International Conference on , pp.1-2, 10-14 Jan. 2009
[16] G. Zhang, J. Jia, T. Wong, H. Bao, Consistent Depth Maps Recovery from a Video Sequence. IEEE Trans. Pattern Anal. Mach. Intell. Vol. 31, No.6, pp. 974-988 (2009).
[17] C. Cigla, and A. A. Alatan, Temporally consistent dense depth map estimation via Belief Propagation, in Proceedings of 3DTV-CON 2009, 4-6 May 2009, Potsdam, Germany.
[18] Sang-Beom Lee, Yo-Sung Ho; "Multi-view Depth Map Estimation Enhancing Temporal Consistency" in Proceedings of 23rd International Technical Conference on Circuits/Systems; Computers and Communications (ITC-CSCC 2008)
[19] De Silva, D.V.S.X.; Fernando, W.A.C.; Arachchi, H.K.; , "A new mode selection technique for coding Depth maps of 3D video," Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on , vol., no., pp.686-689, 14-19 March 2010
[20] Oliver, J., & Malumbres, M. P, "Low-complexity multiresolution image compression using wavelet lower trees," IEEE Transactions on Circuits and Systems for Video Technology, 16(11), 1437–1444, 2006.
[21] Oliver, J.; Lopez, O.; Martinez-Rach, M.; Malumbres, M.P.; , "A General Frame-by-Frame Wavelet Transform Algorithm for a Three-Dimensional Analysis with Reduced Memory Usage," Image Processing, 2007. ICIP 2007. IEEE International Conference on , vol.1, no., pp.I-469-I-472, Sept. 16 2007-Oct. 19 2007
[22] Z.Wang, A.C.Bovik, H.R. Sheikh, E.P. Simoncelli "Image Quality Assessment: From Error Visibility to Structural Similarity" IEEE Transactions on Image Processing, vol. 13, no.4, April 2004
[23] Xinbo Gao, Wen Lu, Dacheng Tao, and Xuelong Li, "Image quality assessment based on multiscale geometric analysis," IEEE Transactions on Image Processing, vol. 18, no. 7, pp. 1409–1423, 2009.
[24] M. Martinez-Rach, O. Lopez, P. Piñol, J. Oliver, and M. Malumbres, "A study of objective quality assessment metrics for video codec design and evaluation," in Eight IEEE International Symposium on Multimedia, vol. 1, ISBN 0-7695-2746-9. San Diego, California: IEEE Computer Society, Dec 2006, pp. 517–524.
[25] H. R. Sheikh, M. F. Sabir, and A. C. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," IEEE Transactions on Image Processing, vol. 15, no. 11, pp. 3440– 3451, 2006.
[26] Z. Wang, A. Bovik, H. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," IEEE Transactions on Image Processing, vol. 13, no. 4, 2004.