Comparative Analysis of Occlusion-Filling Techniques in Depth Image-Based Rendering for 3D Videos

Lucio Azzari
Università degli Studi Roma TRE
Via della Vasca Navale,84
00144 Roma, Italia
+390657337061
azzarilucio@libero.it

Federica Battisti
Università degli Studi Roma TRE
Via della Vasca Navale,84
00144 Roma, Italia
+390657337061
federica.battisti@uniroma3.it

Atanas Gotchev
Tampere University of Technology
P.O.Box 553
33101 Tampere, Finland
+358331154349
atanas.gotchev@tut.fi

ABSTRACT

Due to the recent success of 3D cinema, 3D video has been gathering considerable interest and delivery of 3D video to home TVs and mobile devices has been actively researched. A broadcast-friendly 3D video delivery format is so-called 'view plus depth', where single video sequence is augmented by dense depth information. Based on it, desired views can be synthesized at the receiver side. While well susceptible for effective compression, this representation presents problems related with occluded areas which become visible in the synthesized views. Different occlusion filling techniques have been developed varying in quality and complexity. This contribution aims at providing an objective and subjective comparison of three such techniques. These include adaptive pre-filtering of depth maps, and one simplified and one high-performance impainting techniques. The latter two have been specifically modified for the case of occlusion filling. We have aimed especially at ranking the performance in the case of mobile 3D imaging. Along with objective comparisons, results from subjective tests are presented as well. Both groups of results demonstrate the superiority of the second impainting technique, which is however also the most computationally expensive one.

Categories and Subject Descriptors

I.4 [**Image Processing and Computer Vision**]: Applications.

General Terms

Algorithms, Performance, Experimentation, Human Factors, Standardization.

Keywords

DIBR, Inpainting, Video Processing, Image Processing.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
Copyright 20XX ACM X-XXXXX-XX-X/XX/XX ...$10.00.

1. INTRODUCTION

After the success of 3D digital cinema, research efforts have been focused on delivery of 3D video for home entertainment and to mobile devices. In its simplest form, 3D video is delivered by two synchronized video sequences representing binocular disparity and targeting the left and right eye correspondingly. An alternative format, known as 'view plus depth' (V+D), represents the 3D moving scene by single video sequence augmented by per-pixel depth information in the form of grey-scale image sequence. In terms of imaging, depth maps are piecewise smooth with clearly delineated borders and grey scales proportional to distances. Thus, they are quite susceptible for compression and attractive for 3D video delivery. Given the depth (range) information in an explicit form, desired perspective views can be synthesized by a technique referred to as Depth-Image-Based-Rendering (DIBR) [10]. In DIBR, stereo video pairs can be created by warping the given view, based on geometric rules and with pixel disparity proportional to their relative depth, taken from the given depth map [13]. The principal drawback of V+D representation is that areas occluded in the main view become visible after virtual view rendering. Thus, the correct form of DIBR has to include image warping followed by filling of disoccluded holes. A number of techniques have been proposed in literature: in [14], an average filter is used to fill the holes, resulting in low perceived quality due to the rubber-sheet artifact [7]. In the same paper a pre-processing of the depth map that aims to reduce the size of the holes after warping is presented. A symmetric Gaussian filter is used to smooth the depth map, and to reduce the high contrast. In [13] Zhang et al. suggest using an asymmetric Gaussian filter in order to reduce the vertical edge artifacts resulting from symmetric smoothing. The resulted warped image contains small holes, of about 1 or 2 pixels easy to fill by an averaging filter. A consequence of these algorithms is that the whole depth map is modified, and this could cause some loss of depth cue after the rendering. To cope with depth loss, Park et al. [9] propose a gradient direction-based filter that smoothes only horizontal edges, causing holes. An alternative approach to hole filling without modifying the depth map is to utilized inpainting techniques. Originally, such types of techniques have been developed to support the restoration of damaged parts of images (e.g. scratches) or even to remove objects from images, while filling the gaps with meaningful texture. Inpainting is quite computationally expensive however attractive from

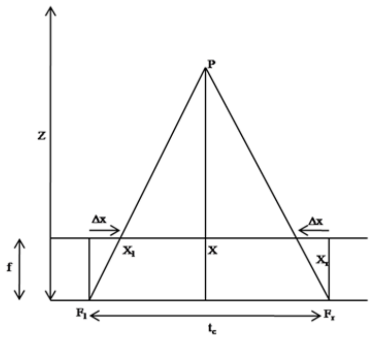


Figure 1: Rendering of the point P.

quality point of view.

This paper aims at comparing three methods for occlusion filling varying in complexity and performance. First method employs pre-processing of the depth-map using the filter presented in [5]. The next two methods do not modify the depth map but rather employ inpainting-based techniques applied directly to the synthesized image. More specifically, the second method is a computationally-efficient inpainting originally proposed by Oliveira et al. [8], which we modified for our occlusion filling needs. The third method is a high-performance and computationally expensive inpaining originally proposed by Criminisi et al. [4], again modified for our needs. The comparisons include both objective and subjective tests.

2. METHODS

2.1 Depth-Image-Based-Rendering

DIBR is a technique that from an 'intermediate image' and a correspondent depth map allows simulating two 'virtual views', one left and one right, and consequently to create a 3D image. As shown in Figure 1, the knowledge of the position of the point P in the original scene (its distance from the camera and its coordinates in the intermediate camera), can be used to simulate its projection in other two virtual cameras translated in the horizontal direction of the distance $\pm t_c/2$, and positioned in the focal points F_l and F_r. The shift in the left and right camera is:

$$X_{l/r} = X + \Delta x_{l/r} \quad (1)$$

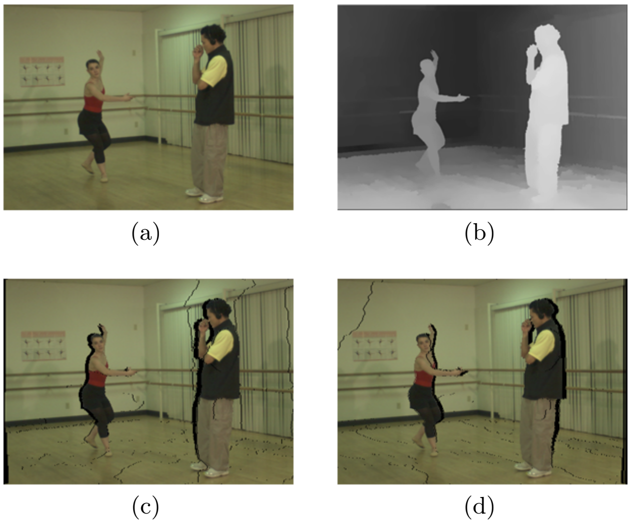
in which $\Delta x_{l/r}$ is the horizontal shift, and it is equal to:

$$\Delta x_{l/r} = \begin{cases} \frac{t_c f}{2Z} & \textit{left view} \\ -\frac{t_c f}{2Z} & \textit{right view} \end{cases} \quad (2)$$

where Z is the depth value of the pixel in the intermediate image, f is the focal length of the camera and $X_{l/r}$ is the resulting horizontal coordinate of the pixel in the left/right virtual camera [13], [6].

To obtain an image with a better quality and to obtain a more realistic depth cue, in [7] the so called 'zero parallax setting' (ZPS) is introduced; it represents the depth plane in which the disparity is zero. This effect is obtained by the simulation of a virtual shift of the sensors, according to [7] and [12], proportionally to the plan that we want to select as ZPS. Equation 1 becomes:

$$X_{l/r} = X + \Delta x_{l/r} + h \quad (3)$$



(a) (b)

(c) (d)

Figure 2: Original frame (a) from the ballet sequence with its relative depth-map (b), and the left (c) and right (d) rendered views.

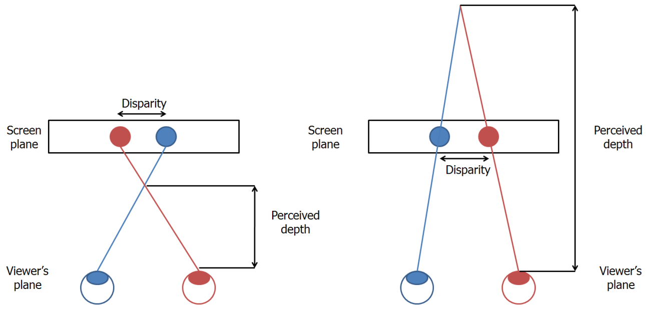


Figure 3: Depth perceived due to ZPS.

where the sensors' shift h is:

$$h = \begin{cases} -\frac{t_c f}{2Z_c} & \textit{left view} \\ \frac{t_c f}{2Z_c} & \textit{right view} \end{cases} \quad (4)$$

In this formula Z_c is the 'convergence plane', and usually it is set to the intermediate distance perceived. By using this method it is possible to obtain a more accurate depth cue as shown in Figure 3. Usually, before warping, a preprocessing of depth map is used to reduce the number of disoccluded areas, and in particular in [9] a smoothing filter that modifies only the horizontal edges is used. In Figure 4 there is the typical flow-chart of a DIBR process.

3. OCCLUSION FILLING

As already stated, the drawback of DIBR techniques is in the filling of the disoccluded regions. Such regions, marked by black color, are shown in Figure 2. The goal of occlusion

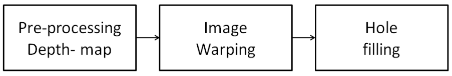


Figure 4: DIBR's flow-chart.

a	b	a
b	0	b
a	b	a

c	c	c
c	0	c
c	c	c

Figure 5: Two different kinds of diffusion kernel.

filling techniques is to fill the holes and to make the resulting image looking 'natural'. Inpainting is particularly suitable technique for occlusion filling. Digital inpainting has been introduced for the first time in [2]. In this work, Bertalmio et al. have proposed to use partial differential equations (PDE) of fluid dynamic to introduce color diffusion like a 'fluid flow' along the so-called isophotes, that are the directions of the color lines inside the holes. The filled parts present a continuation of edges inside the unknown areas. By this approach, textures are not well reproduced due the use of the Laplacian derivative in the diffusion equation. Chan et al. in [3] have used the curvature of isophotes to ensure the continuity of the lines inside the hole. In general, inpainting methods differ in terms of complexity and execution time.

For the kind of holes, resulting from disocclusion, i.e. thin and long, such as those in Figure 2, we selected a modified version of the Oliveira's method [8], which offers doable performance for tolerable computational time. By that approach, the filling process is performed by iterating a convolution of the hole's pixels with one of the diffusion kernel presented in Figure 5. This approach results in regions filled and smoothed similarly to the Bertalmio's algorithm but with reduced complexity. In the original Oliveira's technique, the diffusion of the color inside the hole is obtained using all surrounding pixels. In the case of occlusions, the respective regions are part of the background and only propagation of the background texture is required, therefore we have modified the algorithm accordingly.

The last filling algorithm implemented is the Criminisi's exemplar-based method [4]. Following this algorithm, the region of holes is filled using the surrounding information based on similarity. A window is centered on the boundary of the hole and similar windows (patches) are searched based on the available part of the reference window (pixels outside the hole). A non-local mean of the N most similar patches is formed to fill the missing pixels [11], see Figure 6 for an example. The original algorithm explores the concept of 'priority map', i.e. the way similar patches are prioritized for weighting. The standard method offers the possibility that the first target regions to be filled are taken near the foreground, and consequently the best matches are 'more similar' to the foreground than to the background. In our modification, the information about the depth is exploited in order to create an appropriate priority map and to be able to fill first the regions close to the background, then the foreground. Furthermore, we have extended thee search for similar block toward neighboring previous and following video frames in order to take into account the possible motion of objects along successive frames that would reveal regions which were previously occluded.

4. EXPERIMENTAL RESULTS

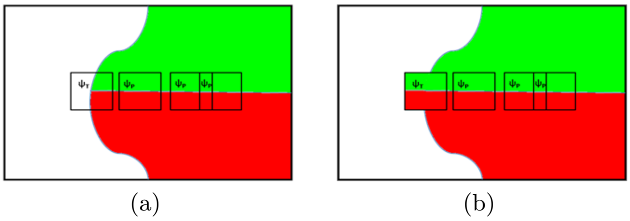


Figure 6: Non-local mean of best matching patches (a) and filling result (b).

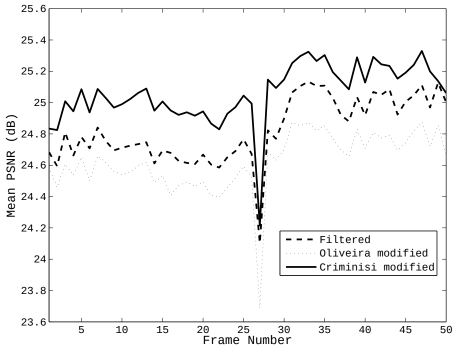


Figure 7: PSNR values for 50 frames.

4.1 Parameter setting

In the performed tests, sequences with suitable resolution for mobile devices (240 × 432 pixels) have been used: the Bullinger, Butterfly, Horse, Car and Quest sequences from Fraunhofer Heinrich Hertz Institute's (HHI) dataset. They present indoor and outdoor scenes, and motion and no-motion camera. Starting from the left views and using the Depth-maps, the right views have been estimated in order to compare them with the originals. For what concerns the algorithm that smoothes the depth-map, the maximum number of filtering iteration has been set to 15, and the parameter h to 50. For Oliveira's method the maximum number of iteration has been fixed to 100 and for Criminisi's technique a 7x7 pixels patch in a research window of 40x40 pixels in the spatial domain, and 11 frames in the temporal domain have been used.

4.2 Objective results

In order to analyze the similarity between the rendered images and the original ones, the Peak Signal-to-Noise Ratio (PSNR) and the Weighted Peak Signal-to-Noise Ratio (WPSNR) among the different warped images and the 'real' right view have been considered. In each video, 50 frames are considered, and for each rendered frame, the PSNR and the WPSNR are computed with the original right view and their values are reported in Figure 7 and 8; finally an average of the values relative to the n-th frame of every video is calculated. Figure 6 show the results of these tests; the first notable aspect is that in each sequence the PSNR and WPSNR values of Park's method are below the Criminisi one; this happens because the first algorithm modifies the depth-map, and consequently the warped image is different from the original; for this reason the perceived depth could be different from the original one.

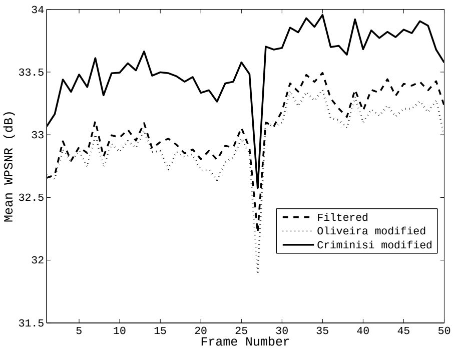


Figure 8: WPSNR values for 50 frames.

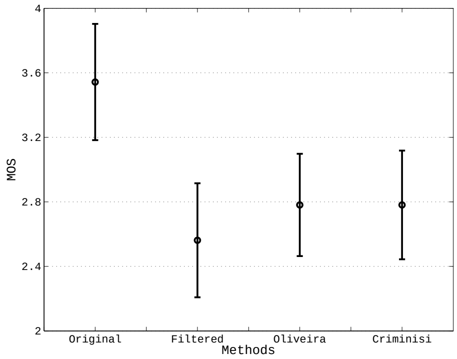


Figure 9: MOS relative to the general video quality.

However the use of these metrics is not sufficient to have a good estimation of the quality of these methods, and for this reason a subjective test has been performed, in order to have a further validation of the validity of the proposed methods.

4.3 Subjective test description

The subjective tests have been performed using the rendered sequences. 21 subjects have undergone the tests, and each of them rated 20 videos (5 original and 15 rendered videos obtained using the three methods under analysis). In particular the test consists in showing to the subject the stereo videos, including the originals, in random order, to have an independent rate of all the sequences as detailed in *ITU - R_B T.500 - 11* [1]. After each sequence, six questions have been asked to the viewer, concerning:

- general video quality;
- objects' borders;
- depth cue;
- annoyance due to flicker artifacts;
- annoyance due to unnatural shape of objects;
- eyes' annoyance.

For the first three questions the rate is a number between 1 and 5, where 1 represents "bad" and 5 is "excellent", while for the last three questions the rate is a number between 0 and 5, where 0 represents the "not annoyance feeling", while 5 is "the very annoying feeling".

4.4 Subjective test evaluation

In this section the Mean Opinion Score (MOS) regarding the general quality of the videos, the depth quality perceived, the annoyance due to the flicker effect and the eyes' annoyance will be reported. Considering the behavior obtained by considering the first question in Figure 9, it is evident that the MOS of the proposed methods are better than the algorithm that filters the depth-map for 0.2 points.

Analyzing the MOS of the annoyance questions it is possible to better explain these results. The general annoyance felt by the viewer, expressed in Figure 10, is very similar for all analyzed methods, but in Figure 11 it is possible to notice that the flicker effect is very annoying principally in the Criminisi modified algorithm; consequently by solving the flickering problem, it is possible to obtain increased MOS values.

It is important to mention that there are two main causes that can generate flicker artifacts:

- the depth map that is used is non-ideal: the correspondence between the pixels in the original image and their corresponding depth values is never exact; in particular if a pixel in the background or in the foreground has a wrong depth value, it could be incorrectly shifted after the rendering; consequently all wrong shifted pixels create a flicker effect during the reproduction of the rendered sequence.

- consecutive frames can generate different filling results: the filling of a single frame, independently from the others, can produce results only apparently correct; in fact, when it is inserted as frame in a video it can appear annoying for the viewers, that are sensitive to the difference between consecutive frames.

To reduce the above-mentioned effects it is necessary to separately operate on them.

In order to improve the accuracy of the depth map, it is possible to process it with a bilateral Gaussian filtering map; in this way the errors made in the estimation of the depth map can be corrected.

To reduce the problems caused by the second effect, it is possible to opportunely tune the weights of the patches in the non-local mean computation; in this way the filling of similar holes in consecutive frames could give similar results reducing the mentioned artifacts. The authors are already working in the application of these enhancements and the first achieved results show that the rendered videos present an increased quality.

Even if the Criminisi modified algorithm has worst performance for the flicker artifacts, the MOS relative to the depth cue perceived by the viewer is the highest among the rendering methods, as shown in Figure 12. This means that the 3D effect is well reproduced using this method.

Finally, a comparative analysis of the computational cost required by the three algorithms has been estimated.

In more details, the method based on the filtering of the depth map requires a number of operations per pixel equal to:

$$(N \cdot N - x) + 1 \tag{5}$$

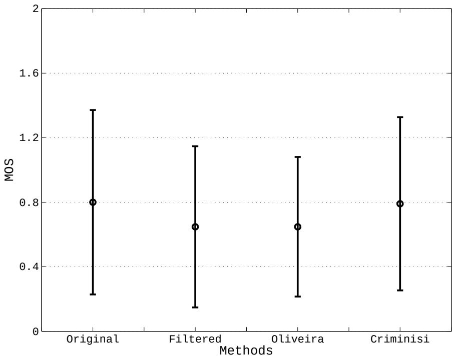


Figure 10: MOS relative to the general eyes' annoyance feeling.

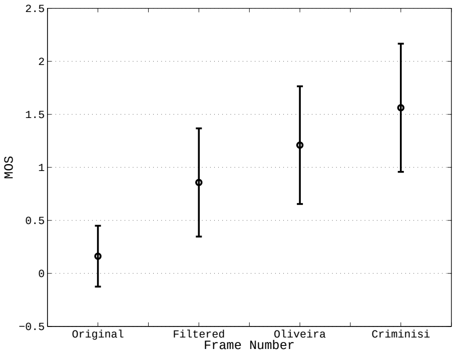


Figure 11: MOS relative to the flickering annoyance.

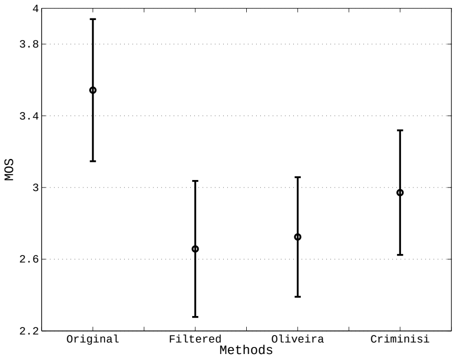


Figure 12: MOS relative to the perceived depth cue.

where N is the dimension of the average filter and x is the number of unknown pixels in the window.

In this algorithm $(N \cdot N - x)$ sums are followed by a division.

The analysis of Oliveira's method resulted in a number of operations per pixel equal to:

$$[(f_{size} \cdot f_{size}) + 1] \cdot T \tag{6}$$

where f_{size} is the dimension of the filter (3 in the proposed scheme), as shown in Figure 5, and T is the number of convolution iterations. In this case nine multiplications plus a sum are iterated T times in order to fill a single pixel.

Finally, in Criminisi's modified algorithm, the number of operations required per pixel is equal to:

$$\frac{\sum_{i=1}^{F} [(M - w_{size} + 1)^2 \cdot (w_{size} \cdot w_{size} - x + 1)] \cdot \gamma_i}{x} \tag{7}$$

where M is the size of the search window, w_{size} is the size of the target window, x is the number of unknown pixels in the window, and γ_i is the part of the search window that contains well-known pixels:

$$\gamma_i = \frac{M \cdot M - X_i}{M \cdot M} \tag{8}$$

where X_i is the number of unknown pixels in the search window. It is important to underline that the distance has been evaluated by means of L1 norm in order to reduce the number of operations to be computed. Eq. 8 is an approximative estimation of the number of iteration per pixel, in fact this number changes for every target window.
From the performed analysis it can be noticed that the fastest algorithm is the first one, in which, in order to fill the disoccluded areas, the average of the surrounding well-known pixels is performed once; in the second presented approach, a convolution has to be iterated T times, thus increasing the time needed for filling a single pixel. The last algorithm allows to fill more than one pixel simultaneously, but the block matching phase slows down the process, thus resulting in a slow method. A possible solution that could speed up the modified Criminisi's method is to find a threshold, defined as a maximum distance value, and to stop the searching algorithm when a sufficient number of similar patches below the selected threshold are found.

5. CONCLUSION

In this paper it is proposed a new approach for the hole filling in DIBR context. In particular two modified inpainting algorithms, that present different computational time, are used in order to fill the disoccluded regions created after the warping of the intermediate image. Their performances are compared with a State-of -the-art method of DIBR using objective and subjective metrics. The PSNR and WPSNR between the rendered and the original videos are computed; these tests show that the best behavior is obtained by Criminisi's modified technique. At the same time a subjective test for 3D quality has been created in order to analyze the perceived quality of the rendered videos. The results of this test show that the best quality is achieved by Criminisi's algorithm, but its performances have to be improved in order to have a real application in video stereo streaming; in particular it is necessary to reduce the flickering effect in the rendered image, that is annoying for a great part of viewers, and consequently obtain a better subjective rate. In general Criminisi's method presents better results with respect to the other algorithms, and this is encouraging for future improvement. At the same time its performances should be improved in order to adapt it to real application.

6. REFERENCES

[1] Methodology for the subjective assessment of the quality of television pictures. *RECOMMENDATION ITU-R BT.500-11.*

[2] M. Bertalmio, G. Sapiro, V. Caselles, and C. Ballester. Image Inpainting. *Proc. 27th annual conference on Computer graphics and interactive techniques*, pages 417 – 424, 2000.

[3] T. F. Chan and J. Shen. Nontexture Inpainting by Curvature-Driven Diffusions. *Journal of Visual Communication and Image Representation*, 12:436–449, 2001.

[4] A. Criminisi, P. Perez, and K. Toyama. Object Removal by Exemplar-Based Inpainting. *IEEE Transactions on Image Processing*, 13, 2004.

[5] C. Fehn. Depth-Image-Based Rendering (DIBR), Compression and Transmission for a New Approach on 3D-TV. *Proc. SPIE Stereoscopic Displays and Virtual Reality Systems XI*, pages 93–104, 2004.

[6] K. H. Jung, Y. K. Park, J. K. Kim, H. Lee, K. J. Yun, N. H. Hur, and J. W. Kim. 2D/3D Mixed Service in T-DMB System Using Depth-Image Based Rendering. *Proc. 10th International Conference on Advanced Communication Technology (ICACT 2008)*, 3:1868 – 1871, 2008.

[7] L. Lipton. *Foundations of the Stereoscopic Cinema - A Study in Depth.* Van Nostrand Reinhold, New York, NY, USA, 1982.

[8] M. M. Oliveira, B. Bowen, and R. McKenna. Fast Digital Image Inpainting. *Proc. International Conference on Visualization, Imaging and Image Processing (VIIP2001)*, pages 261–266, September 2001.

[9] Y. K. Park, K. Jung, Y. Oh, S. Lee, J. K. Kim, G. Lee, H. Lee, K. Yun, N. Hur, and J. Kim. Depth-image-based rendering for 3DTV service over T-DMB. *Signal Processing: Image Communication*, 24:122–136, 2009.

[10] A. Redert, M. O. D. Beeck, C. Fehn, W. Jsselsteijn, M. Pollefeys, L. V. Gool, E. Ofek, I. Sexton, and P. Surman. ATTEST: Advanced Three-dimensional Television System Technologies. *Proc. 1st International Symposium on 3D Data Processing Visualization and Transmission (3DPVT 2002)*, 2002.

[11] A. Wong and J. Orchard. A non local-means approach to exemplar-based inpainting. *Proc. IEEE International Conference on Image Processing (ICIP 08)*, pages 2600–2603, 2008.

[12] A. Woods, T. Docherty, and R. Koch. Image Distortions in Stereoscopic Video Systems. *Proc. SPIE Stereoscopic Displays and Applications IV*, pages 36–48, 1993.

[13] L. Zhang and W. J. Tam. Stereoscopic image generation based on depth images for 3D TV. *IEEE Transactions on Broadcasting*, 51:191–199, 2005.

[14] L. Zhang, W. J. Tam, and D. Wang. Stereoscopic image generation based on depth images. In *Proc. IEEE International Conference on Image Processing (ICIP 04)*, pages 2993–2996, 2004.