Pattern Recognition Letters

Authorship Confirmation

Please save a copy of this file, complete and upload as the "Confirmation of Authorship" file.

As corresponding author I, _____, hereby confirm on behalf of all authors that:

1. This manuscript, or a large part of it, has not been published, was not, and is not being submitted to any other journal.
2. If presented at or submitted to or published at a conference(s), the conference(s) is (are) identified and substantial justification for re-publication is presented below. A copy of conference paper(s) is(are) uploaded with the manuscript.
3. If the manuscript appears as a preprint anywhere on the web, e.g. arXiv, etc., it is identified below. The preprint should include a statement that the paper is under consideration at Pattern Recognition Letters.
4. All text and graphics, except for those marked with sources, are original works of the authors, and all necessary permissions for publication were secured prior to submission of the manuscript.
5. All authors each made a significant contribution to the research reported and have read and approved the submitted manuscript.

Signature_____ Date_____

List any pre-prints:

Relevant Conference publication(s) (submitted, accepted, or published):

Justification for re-publication:

Graphical Abstract (Optional)

To create your abstract, please type over the instructions in the template box below. Fonts or abstract dimensions should not be changed or altered.

Type the title of your article here

Author's names here

This is the dummy text for graphical abstract. This is the dummy text for graphical abstract.

Pattern Recognition Letters
journal homepage: www.elsevier.com

Occlusion filling for high quality multiview synthesis

Geetha Ramachandran[a,**], Markus Rupp[a], Walter Kropatsch[b]

[a]*Institute of Telecommunications, Vienna University of Technology, Gusshausstrasse 25/389, 1040 Vienna, Austria*
[b]*Institute of Computer Graphics and Algorithms, Vienna University of Technology, Favoritenstrae 9/186, 1040 Vienna, Austria*

ABSTRACT

We investigate the problem of generation of new images and their corresponding disparity maps at vantage positions intermediate to a set of given stereo views. Linear interpolation is applied to generate the candidate views and disparity maps. Based on the input images, these candidate views contain a varying degree of occlusion. We present and compare two separate methods to fill these holes. We evaluate our methods using the Middlebury stereo dataset. Through our experiments, we show that both the methods we propose produce high quality images with higher PSNR and SSIM compared to state-of-the-art methods.

1. Introduction

For a fully immersive 3D viewing experience, autostereoscopic 3D displays offer an ideal solution. They enable multiple viewers to perceive 3D without the burden of external wearables. These displays are typically grouped into two-view, two or multiview with fixed viewing zones or head/pupil tracked and super multiview, Dodgson (2005), Urey et al. (2011). We focus here on multiview displays. These systems work with the input and output of multiple streams of content. This problem has long been a matter of interest, one of the earliest works of literature on the topic being by Lippmann (1908). State-of-the-art displays incorporate optical elements such as lenticular arrays or parallax barriers in front of, for instance, a liquid-crystal display panel in order to send the image information to distinct directions the so-called 3D viewing zones. This is facilitated through the transmission of multiple stereo views. Here, we explore methods to reduce the number of stereo data streams being simultaneously transmitted.

The key idea we investigate is the generation of views from new vantage points. Specifically, we consider the problem of interpolation of stereo views, given their corresponding disparity maps, to produce multiple intermediate views. For each new vantage point, a new image and its corresponding new disparity map are to be generated. The main challenge here is the generation of occlusion (hole) free views. There are mainly two reasons for the occurrence of occlusions. First, new positions may appear in the view that were occluded in the original stereo views. Secondly, these may be a by-product of lack of disparity information in the original disparity maps for instance due to the illumination conditions in the original images. We elaborate on two methods to generate occlusion-free images.

To generate the initial intermediate views, we use linear interpolation. This creates a primary image and disparity map at the new vantage point. To hide the occlusions in the disparity map, the variance of the disparity values surrounding each occlusion is computed, based on which a foreground or a background value is assigned. For the color image, first we convert it from the RGB color space to the CIELAB (LAB) color space, Hunter (1948). The LAB color space separates out the luminance from the chrominance and is perceptually uniform. We then apply the variance check on the LAB layers separately. This method, though performs better during evaluation for some datasets (as shown in Section 4.2), is not so successful with images with textured backgrounds. To improve this, we replace the holes with patches from the original input images.

The key contribution of our work is the demonstration of the application of patch based methods for occlusion filling in view synthesis. Also, it is simple to implement with only two main steps, but provides higher peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) values. The PSNR values measure the quality of the reconstructed image, while the SSIM gives a good indication of the similarity between the images.

**Corresponding author: Tel.: +43 58801 38906;
e-mail: geetha.ramachandran@tuwien.ac.at (Geetha Ramachandran)

We evaluate our methods on the Middlebury Stereo Database Scharstein and Pal (2007).

The rest of the paper is structured as follows: In Section 2, we look at similar recent work in the area of view synthesis. We describe the generation of the new view and disparity map in Section 3 and the method to fill out the occlusions in Section 4. This is followed by the experimental evaluation in Section 5 and our findings and conclusions in Section 6.

2. Related literature

One of the early works on the topic of view synthesis applies image morphing to adjacent images to create a new image for an in-between viewpoint, Chen and Williams (1993). Here, the camera transformation and image range data is used to automatically determine the correspondence between two or more images. Linear interpolation is applied to generate new pixel coordinates for the new viewpoint as a replacement for the coordinates given by the perspective viewing matrix. Holes are filled out using the background colors or by using more source images to generate the view.

A seminal work on the topic of view interpolation in videos is described by Zitnick et al. (2004). To capture the data, eight cameras are set up around a dynamic scene spanning a $1D$ arc of about 30°. Each image is first smoothed and then segmented based on neighboring color values. An initial disparity space distribution is computed for each segment in each camera by matching points in neighboring images. The disparities are refined by projecting the pixels from one image to the next to check for consistency. Typically, pixels that occur along the border boundary of objects will receive contributions from both the foreground and background. This is called the mixed pixel problem. Using the original values during image-based rendering will result in visible artifacts. This is resolved by computing matting information within a neighborhood of four pixels from all depth discontinuities. Within these neighborhoods, foreground and background colors along with opacities (alpha values) are computed using Bayesian image matting. Similar work using multiple input images for virtual viewpoint replay in soccer games is investigated by Inamoto and Saito (2007) where the background and correspondences among the images are found manually and projective geometry is applied for interpolation of the views.

As an improvement to the work by Zitnick et al. (2004), image based 3D warping is employed by Smolic et al. (2008). Layer separation is performed by looking for depth discontinuities. Then, samples of the original 2D views are projected into 3D space and forward projected into an intermediate view. Holes are filled using median filters or by simply applying background disparities. The work in Manap and Soraghan (2011) uses a similar approach and investigates the separation of the depth map into multiple layers using an image histogram distribution. The novel view is synthesised at each layer independently. The pixel interpolation is performed by masking each particular layer and finally blending the layers together.

Jain et al. (2011) follow a four step process towards view interpolation. The initial view is first generated using interpolation. The generated view and disparity map and image are then refined by applying morphological operations. Occlusions are filled in the disparity map by looking at the variance in windows around the missing pixels. For the image, in each hole region, a k-means segmentation is applied to get color clusters. The color is assigned based on the smallest median distance to each cluster. This paper provides numerical results for the new views in terms of peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM) values for experiments performed on the Middlebury Stereo Database. Ramachandran and Rupp (2012) reduce the complexity of this method by introducing non-integer interpolation and simplifying the occlusion filling in color images by only considering whether the hole occurs in the foreground or in the background. Our algorithm is an improvement over these works and we provide our improved PSNR and SSIM values in Section 5.

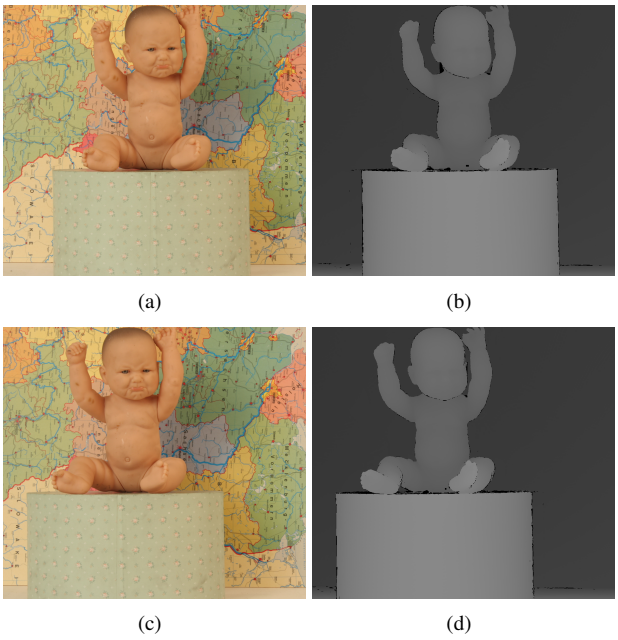


Fig. 1. 'Baby1' dataset from Middlebury Stereo Database. (a)-(b) Left stereo view and disparity map (c)-(d) Right stereo view and disparity map

3. Initial view generation

Given a set of stereo views (S_L, S_R) and their corresponding disparity maps (D_L, D_R) with M rows and N columns, here shown in Figure 1, two candidate intermediate views and disparity maps are generated. Since we are taking into account rectified views here, the displacement of the pixels from the left to the right stereo view are in a horizontal line. The position of the virtual camera for the new vantage point is at α such that $0 < \alpha < 1$. As a first step, we apply linear interpolation to the images and disparity maps and generate a set of candidate images. These are improved by applying bilinear interpolation to points where the displacement of the pixels are not integer valued. For more details on this, please refer to the work by Ramachandran and Rupp (2012). The equations below describe the method for the left image for each point (m, n):

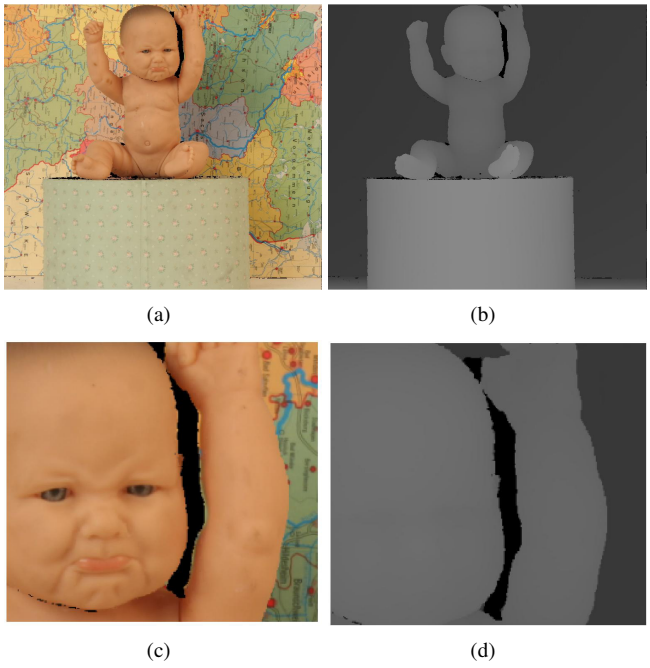


Fig. 2. Results from Section 3(a) Generated intermediate view (b) Disparity map of the intermediate view (c) - (d)Zoomed in occlusion in view and disparity map

$$\delta = \lceil n - \alpha D_L(m,n) \rceil - [n - \alpha D_L(m,n)], \quad (1)$$

$$T_L(m, \lceil n - \alpha D_L(m,n) \rceil) = S_L(m,n), \quad (2)$$

$$T_L(m, \lfloor n - \alpha D_L(m,n) \rfloor) = \frac{S_L(m,n) - \delta T_L(m, \lceil n - \alpha D_L(m,n) \rceil)}{1-\delta}. \quad (3)$$

And for the right candidate image,

$$\delta = \lceil N - n + (1-\alpha) D_R(m, N-n) \rceil - [N - n + (1-\alpha) D_R(m, N-n)], \quad (4)$$

$$T_R(m, \lceil N - n + (1-\alpha) D_R(m, N-n) \rceil) = S_R(m, N-n), \quad (5)$$

$$T_R(m, \lfloor N - n + (1-\alpha) D_R(m, N-n) \rfloor) = \frac{S_R(m, N-n) - \delta T_R(m, \lceil N-n + (1-\alpha) D_R(m, N-n) \rceil)}{1-\delta}. \quad (6)$$

where δ denotes the non-integer factor. The candidate views (T_L, T_R) are then merged as in Jain et al. (2011). The mixed pixel problem is solved by retaining pixels with greater disparity as these indicate foreground values and hence are not likely to be occluded. The intermediate image and disparity map for the dataset Baby1 are shown in Figure 2.

4. Occlusion filling

From Figure 2, it is seen that the procedure described in Section 3 result in the occurrence of occlusions (for instance, the black section between the baby's head and left arm, shown in) in the generated images and disparity maps. We describe below methods to place pixels into these holes.

4.1. Occlusions in disparity maps

We modify the method described in Jain et al. (2011) to fill out the occlusions in the disparity map. As a first step, the existing disparity values are normalized to lie in the range $[0, 1]$. This is necessary so that the cost function to be used later gives a more accurate evaluation. We apply normalization using feature scaling for this:

$$NDM = \frac{DM - min(DM)}{max(DM) - min(DM)} \quad (7)$$

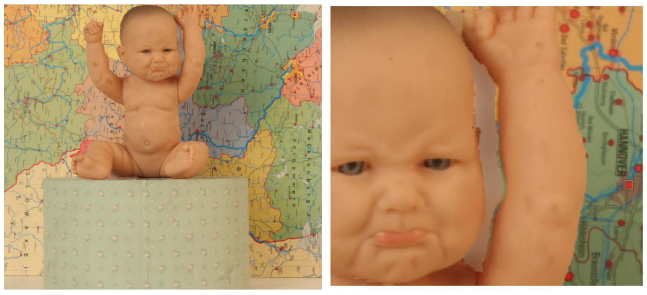
Next, a window w with $N \times N$ block of pixels is chosen around each hole. In addition to the central hole, this block may contain actual disparity information or may be filled only with other holes. A minimal threshold is set such that the block is expanded till it accommodates disparity information. A histogram with B bins and the variance σ_w^2 of the available disparity information for the window w is calculated. The following cost function is then computed:

$$l(b_i) = \beta \sigma_w^2 b_i + \frac{1}{c(b_i)}, \; 1 \le i \le B, \quad (8)$$

where b_i is the ith bin of the histogram, β is a tuning parameter and $c(b_i)$ is the number of elements in the bin b_i. If the first term of the cost function in Equation (8) is higher than the second, it is an indication that the overall disparity variation is low in w. Hence, the lowest disparity of the neighboring values indicating the background is chosen to be assigned to the hole. Otherwise, if the second term of the cost function exceeds the first, the mode is chosen which is the most commonly occurring disparity value. For our experiments, we empirically set the values to be $N = 120$, $B = 10$ and $\beta = 1000$. The value of N is suited to cover most hole sizes that might occur in an image of size 1110×1240, as in the Middlebury dataset. The term β is chosen in such a way that the variance σ_w^2 does not significantly reduce the first term of the cost function. If initially more than 75% of w are holes, then w is expanded to be of size $(N + \Delta) \times (N + \Delta)$ where we choose Δ to be 12. The results of this step are shown in Figure 4(d).

4.2. Occlusions in color images

To fill in the holes in the color images, as a first step, the image is changed from the RGB to the LAB color space. Separating out the luminance and chrominance helps to segregate the layers with brightness and shadow information (L) versus the layers with color information (A, B).



(a) (b)

Fig. 3. Results from Section 4.2.1(a) Generated view using method (b) Zoomed in occlusion region

4.2.1. Variance based occlusion filling

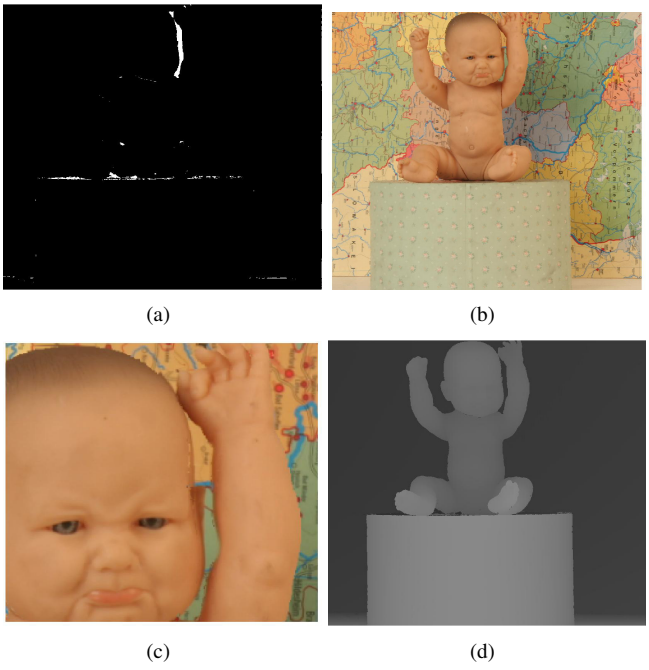
Each component of the LAB image is separately processed. The basic steps of the occlusion filling process remain the same as that for the disparity map. The intensity values are normalized and the same cost function is applied. The key difference is in how the values are chosen to fill the holes. For each hole x_{mn}, when the first term of the cost function is higher than the second, the average of the neighboring pixels from the surrounding window w is set. Here, since we are dealing with colors, the minimal value makes no sense. Otherwise, when the second term of the cost function is higher, as with the disparity map, the mode is chosen. The result is shown in Figure 3(a).

As can be seen from Table 1, the numerical values show an improvement over the work by Jain et al. (2011). This method proves effective in instances where the occlusions are small or when the occluded regions are not heavily textured. However, as seen in Figure 3(b), using the method above for filling in occlusions results in washed over patches of color. All information about the texture is lost. In instances where stereo images have shadowy regions, bigger occlusions occur in the disparity maps. Correspondingly, using the available disparity information for creation of new views creates large occlusions in the synthesized views. Application of this method does not prove effective in these instances. We fix this with the method in the following section.

4.2.2. Filling hole patches

As opposed to filling each hole individually, here, we look at the filling of holes patch wise. To this end, as a first step, we convert the L component into a binary image, Figure 4(a). For each hole x_{mn}, the eight adjacent neighbours are examined. Iteratively, this process is continued until a patch P is identified. We set a threshold to the size of the patches that are considered here. When the patch size is very small, the quality of the image is not largely affected by the lack of texture information. This is demonstrated in Section 5.

For each P, the corresponding disparities from the displaced initial views are taken into account. We apply the same principle of taking the minimal of these disparities for each point in the patch. These values are applied to identify the corresponding patches (P_L, P_R) in the candidate views (T_L, T_R). As seen in Section 3, there are holes (x_{P_L}, x_{P_R}) in both these views. We use this information to choose from which patch the colors are



(a) (b)

(c) (d)

Fig. 4. Results from Section 4.2.1(a) Binary generated image with holes (b) Image with occlusions filled out (c) Zoomed in occlusion region (d) Generated disparity map from Section 4.1

filled as shown in Equation (9).

$$P = \begin{cases} P_{AL} & ; \#x_{P_L} < \#x_{P_R} \\ P_{AR} & ; \#x_{P_R} < \#x_{P_L} \end{cases} \tag{9}$$

where # indicates the number of non-hole pixels present in the patch. We choose the patch from the candidate view which has lesser number of holes. This patch is then placed in the hole. This process is repeated until all the holes have been filled up. The final result is shown in Figure 4(b). Figure 4(c) shows that the same occlusion patch seen in Section 4.2.1 has been filled with the background texture information.

5. Experimental evaluation

The proposed method was tested on the 27 datasets from the Middlebury Stereo Database. These datasets contain seven views of the same scene and disparity maps of the views 1 and 5. Using view 1 and view 5 and the corresponding disparity maps, the intermediate views and disparity maps for the positions in between, views 2, 3 and 4, with $\alpha = 0.25$, $\alpha = 0.5$ and $\alpha = 0.75$ respectively, were generated. These values are indicated in Table 2.

Experiments were conducted to investigate the best threshold for the size of the patch. It was observed over multiple datasets that the PSNR values consistently went up with decreasing patch size. The results are shown in Figure 5.

We compare the PSNR and SSIM values of the generated views to the results from Jain et al. (2011) which are described at `http://videoprocessing.ucsd.edu/~ankitkj/research/viewsynthesis`. The average PSNR and SSIM values are shown in Table 3. The average values of

Middlebury Dataset	Method in Jain et al. (2011)		Method in Ramachandran and Rupp (2012)		Method proposed	
	PSNR(dB)	SSIM	PSNR(dB)	SSIM	PSNR(dB)	SSIM
Aloe	28.79	0.919	27.69	0.884	29.29	0.922
Baby1	33.21	0.953	32.95	0.941	34.24	0.948
Cloth1	35.00	0.965	33.37	0.937	36.06	0.971
Lampshade1	31.60	0.961	31.15	0.959	32.00	0.962
Moebius	33.35	0.946	32.72	0.931	33.83	0.939
Plastic	37.91	0.983	37.99	0.980	37.93	0.976

Table 1. Comparison of PSNR and SSIM values of variance based occlusion filled images with method by Jain et al. (2011) and Ramachandran and Rupp (2012) for $\alpha = 0.5$.

Middlebury Dataset	PSNR(dB)			SSIM		
	View 2	View 3	View 4	View 2	View 3	View 4
Aloe	29.52	29.64	29.69	0.920	0.925	0.916
Art	31.95	30.42	31.76	0.938	0.935	0.936
Baby1	33.57	34.45	34.17	0.941	0.950	0.943
Books	29.95	30.65	30.51	0.929	0.935	0.937
Cloth1	35.48	36.14	35.20	0.966	0.971	0.964
Dolls	32.70	31.86	32.83	0.944	0.944	0.944
Flowerpots	21.29	25.11	21.33	0.919	0.942	0.923
Lampshade1	34.83	35.26	32.71	0.963	0.966	0.962
Midd1	31.79	32.00	32.37	0.946	0.949	0.947
Moebius	34.07	33.76	34.01	0.936	0.942	0.939
Plastic	36.79	37.82	38.78	0.975	0.977	0.979
Rocks1	28.27	32.61	27.51	0.932	0.939	0.930
Wood1	37.18	38.11	35.14	0.939	0.944	0.929

Table 2. PSNR and SSIM values for views with $\alpha = 0.25$, $\alpha = 0.5$ and $\alpha = 0.75$.

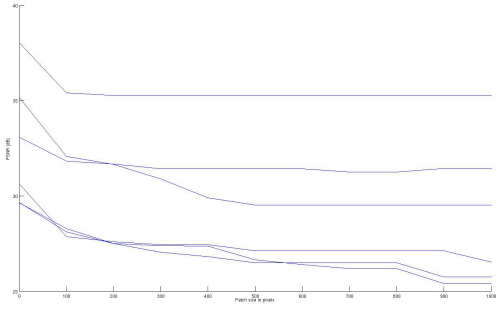


Fig. 5. A plot of the variation of the patch size with the PSNR for different datasets

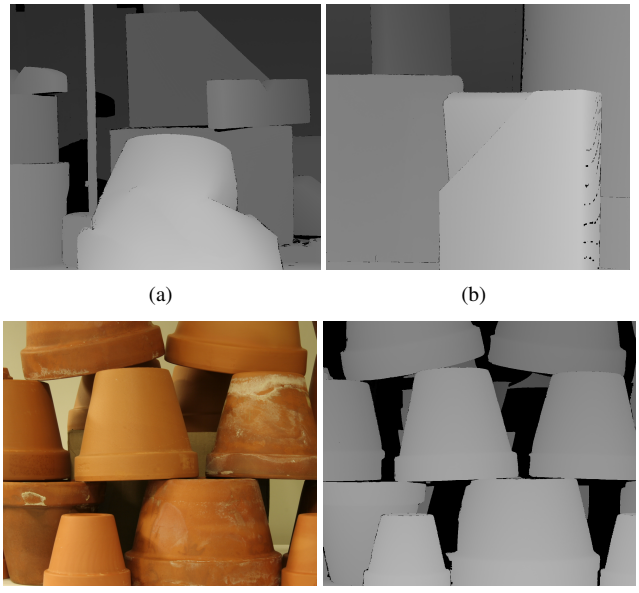


Fig. 6. (a)-(b) Left disparity maps for datasets Lampshade1 and Plastic (c)-(d) Left image and disparity map for dataset Flowerpots

the work by Jain et al. (2011) are slightly higher than those indicated on the website because here we recalculated them logarithmically.

The datasets Lampshade1 and Plastic are very good examples of when the holes in the original disparity maps play a key factor for filling in the occlusions. To illustrate, we show the left disparity maps of these datasets in Figure 6. There are continuous hole regions in the Lampshade1 disparity map, shown by the black region. As can be seen from our values in Table 1 and Table 3, for Plastic, the variance based method provide relatively better PSNR values than the patch based methods, while for Lampshade1, the patch based method proves better. The Plastic dataset is a good example for when this method works best: because the objects have minimal surface texture and the scene is illuminated such that there are minimal shadows.

The methods described above would fail when the original images contain huge patches of shadows, there is a complete lack of disparity information in the original disparity map, and hence the corresponding patch cannot be identified. This is il-

Middlebury Dataset	Method in Jain et al. (2011)		Method proposed	
	PSNR(dB)	SSIM	PSNR(dB)	SSIM
Aloe	28.82	0.920	29.62	0.920
Art	31.69	0.949	31.43	0.936
Baby1	33.59	0.955	34.08	0.945
Books	30.18	0.931	30.38	0.934
Cloth1	35.14	0.965	35.62	0.967
Dolls	31.66	0.949	32.48	0.944
Flowerpots	22.17	0.919	22.97	0.928
Lampshade1	31.55	0.962	34.40	0.963
Midd1	30.80	0.946	32.06	0.947
Moebius	33.46	0.946	33.95	0.939
Plastic	37.98	0.984	37.87	0.977
Rocks1	26.39	0.903	30.08	0.934
Wood1	36.31	0.941	36.98	0.937

Table 3. Comparison of average PSNR and SSIM values with method in Jain et al. (2011).

lustrated in Figure 6 with the dataset Flowerpots. We obtain an average PSNR of 22.97 and SSIM of 0.928 (as compared to 22.17 and 0.919 by Jain et al. (2011)), which is far lower in comparison to other datasets.

The entire code for this work was developed in Matlab. To measure the time it takes, we ran the code on an Ubuntu machine with an Intel i7 processor and 4 GB of RAM, using one core. The average time to run for the 27 datasets was 8 seconds. This is significantly less than the 12 seconds required by Jain et al. (2011).

6. Conclusions

We present here an efficient and effective method for the synthesis of images and disparity maps at new intermediate vantage points to reduce the stream of input data to 3D autostereoscopic displays. We explore two different ways of filling in occlusion regions in the new view. Both these methods prove to have higher PSNR and SSIM values than existing comparable methods. The key consideration in the choice of methods should be how much percentage of the original disparity maps are occluded.

Looking forward, we would like to extend this method to non-rectified video streams. To this end, porting the code to a compiled language would make for realistic run times.

References

Chen, S.E., Williams, L., 1993. View interpolation for image synthesis, in: Proceedings of the 20th Annual Conference on Computer Graphics and Interactive Techniques, ACM, New York, NY, USA. pp. 279–288. URL: http://doi.acm.org/10.1145/166117.166153, doi:10.1145/166117.166153.

Dodgson, N., 2005. Autostereoscopic 3d displays. Computer 38, 31–36. doi:10.1109/MC.2005.252.

Hunter, R., 1948. Minutes of the thirty-first meeting of the board of directors ofthe optical society of america, incorporated. Journal of the Optical Society of America 38, 651–651. URL: http://www.osapublishing.org/abstract.cfm?URI=josa-38-7-651.

Inamoto, N., Saito, H., 2007. Virtual viewpoint replay for a soccer match by view interpolation from multiple cameras. Multimedia, IEEE Transactions on 9, 1155–1166. doi:10.1109/TMM.2007.902832.

Jain, A., Tran, L., Khoshabeh, R., Nguyen, T., 2011. Efficient stereo-to-multiview synthesis, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Prague, Czech Republic. pp. 889–892. doi:10.1109/ICASSP.2011.5946547.

Lippmann, G., 1908. Photographies integrales. Comptes Rendus 146, 446–451.

Manap, N., Soraghan, J., 2011. Novel view synthesis based on depth map layers representation, in: 3DTV Conference: The True Vision - Capture, Transmission and Display of 3D Video (3DTV-CON), 2011, pp. 1–4. doi:10.1109/3DTV.2011.5877181.

Ramachandran, G., Rupp, M., 2012. Multiview synthesis from stereo views, in: Systems, Signals and Image Processing (IWSSIP), 2012 19th International Conference on, pp. 341–345.

Scharstein, D., Pal, C., 2007. Learning Conditional Random Fields for Stereo, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Minneapolis, MN, USA. pp. 1–8. doi:10.1109/CVPR.2007.383191.

Smolic, A., Muller, K., Dix, K., Merkle, P., Kauff, P., Wiegand, T., 2008. Intermediate View Interpolation based on Multiview Video plus Depth for Advanced 3D Video Systems, in: 15th IEEE International Conference on Image Processing (ICIP), San Diego, CA, USA. pp. 2448–2451. doi:10.1109/ICIP.2008.4712288.

Urey, H., Chellappan, K., Erden, E., Surman, P., 2011. State of the Art in Stereoscopic and Autostereoscopic Displays. Proceedings of the IEEE 99, 540–555. doi:10.1109/JPROC.2010.2098351.

Zitnick, C.L., Kang, S.B., Uyttendaele, M., Winder, S., Szeliski, R., 2004. High-quality video view interpolation using a layered representation, in: ACM SIGGRAPH 2004 Papers, ACM, New York, NY, USA. pp. 600–608. URL: http://doi.acm.org/10.1145/1186562.1015766, doi:10.1145/1186562.1015766.