MULTIVIEW SYNTHESIS FROM STEREO VIEWS

Geetha Ramachandran and Markus Rupp

Institute of Telecommunications
Vienna University of Technology
Gusshausstrasse 25/389, 1040 Vienna Austria
{gramacha, mrupp}@nt.tuwien.ac.at

ABSTRACT

A simple and efficient technique for view synthesis from a pair of rectified stereo views is presented here. Depth and colour information from the stereo pair is used to generate an intermediate view and its corresponding disparity map at any given position along the horizontal baseline. Holes in the generated view and disparity map are filled by using the disparity values to separate the hole positions into those present in the foreground or background layers. The results obtained are evaluated using objective quality measures and are compared with other state of the art methods. The high quality intermediate images generated here can be used in applications such as multiview autostereoscopic displays and free-viewpoint television.

Index Terms— view synthesis, disparity map, 3D video, occlusion, multiview

1. INTRODUCTION

The evolution of 3D media stems from the desire for a realistic viewing experience and this in turn has led to great leaps in display technology. Of these, autostereoscopic displays stand out with their capability to allow the viewer to experience 3D without the inconvenience of glasses or other user-mounted devices. Autostereoscopic displays may be two-view, multiview with fixed viewing zones or head/pupil tracked and super multiview [1]. In multiview displays, multiple stereo pairs are presented to the viewer to provide the appearance of a desired motion parallax. To cater to these displays, the same scene needs to be captured from different positions along the viewing field. The challenge in 3D content creation and transmission arises from the difficulty to capture and transfer streams of data from multiple cameras. One of the ways to address this challenge is to generate intermediate views given left and right stereo views. This paper provides a simple and efficient way to generate views at any given position between a left and right view pair.

The process of generating an intermediate view involves obtaining colour and depth information from the given views and disparity maps and mapping this information to the new view. In positions that are occluded in both the left and right views (holes), this is especially challenging. In this paper, we use the disparity information to infer foreground and background positions and fill the holes accordingly.

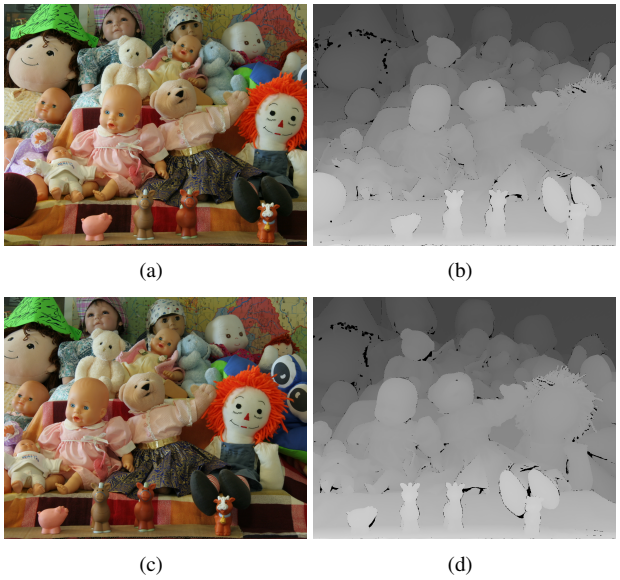
We demonstrate our results using the Middlebury Stereo Database [2]. The paper is organized as follows: a review of the state of the art in view synthesis is provided in Section 2, our methodology is presented in Section 3, the results of our proposed algorithm are presented in Section 4 and Section 5 describes our conclusions and the direction we plan to take in the future.

2. STATE OF THE ART IN VIEW SYNTHESIS

View synthesis using multiple cameras has been explored in [3]. Here, a volumetric model is created from multiple input views to provide correspondences and two views are interpolated using these correspondences to provide a new intermediate view. More recently, a Virtual Video Camera has been proposed to generate free-viewpoint video from unsynchronized, uncalibrated video streams [4].

View synthesis from a stereo pair is used in [5]. Here, the intermediate view is generated from a rectified stereo pair by creating multiple layers of the image based on disparity and then blending them to create a single new image. In [6], foreground and background boundary layers are determined to deal with depth discontinuities and these are fused with a reliable layer to create the final image. Image inpainting aims at providing visual uniformity with the surroundings of occlusions. The paper in [7] describes a technique to fill in the holes of an image and its corresponding disparity map simultaneously by taking depth continuity into account. It is, however, difficult to find numerical results illustrating the efficiency and reliability of these methods and hence a comparison in terms of objective quality is seldom found.

The paper in [8] addresses this problem by providing peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM) values for experiments performed on the Middlebury Stereo Database. The algorithm provided takes 12 seconds per synthesized view and has an average SSIM of 0.95 with



(a) (b) (c) (d)

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

a PSNR of 33.03 dB. The main focus here is on filling in the holes with accurate depth and colour information. However, it fails to take into account the distinction between the background and foreground layers while filling the holes in the generated view and this results in a complicated algorithm. The algorithm we propose provides a simple method to synthesize intermediate views from a rectified stereo pair and we validate our results by preserving high PSNR and SSIM values when compared with those obtained in [8].

3. METHODOLOGY

The proposed algorithm to generate interpolated views between given stereo views uses the following approach. As the first step, the interpolated view and its disparity map are derived by taking into consideration the position at which the new view is to be placed (Section 3.1). Occlusions occur in areas which are hidden in one or both views. In consequent steps, these occlusions are filled in, based on information from neighbouring pixels (Section 3.2-3.3). The sections below describe the steps in detail.

3.1. Interpolating stereo views and disparity maps

Given the stereo views (S_L, S_R) and their corresponding disparity maps (D_L, D_R), as shown in Fig. 1, the first step is to generate two candidate intermediate views from each of the stereo views. The stereo views used here consist of rectified images. Hence, it can be assumed that the correspondences between points in the images occur along horizontal lines. As suggested in [8], a normalized baseline is considered and the left and right cameras are set to be at positions 0 and 1. The position of the virtual camera is at a position α such that $0 < \alpha < 1$. The disparity maps give an indication of the view and illumination disparities between the stereo views. The position of the pixels in the new view is determined by shifting the pixels by scaled disparities. The left and right candidate views T_L and T_R are obtained by displacing pixels by scaling the left disparity map D_L by α and the right disparity map D_R by $1-\alpha$ respectively. Assuming $\alpha D_L(m,n)$ and $(1-\alpha)D_R(m,N-n)$ are integers,

$$T_L(m, n - \alpha D_L(m,n)) \;=\; S_L(m,n) \tag{1}$$

$$T_R(m, N - n + (1-\alpha)D_R(m,N-n)) \;=\; S_R(m,N-n) \tag{2}$$

where $M \times N$ represents the sizes of the image and the disparity map and m and n index the M rows and N columns, respectively.

Typically, the displacements of pixels from frame to frame are not always integer valued. Hence, shifting the pixels by integer values alone introduces inaccuracies in the generated intermediate images. This is a common problem faced in motion compensation. Motion compensated prediction makes use of fractional-pel accuracy in order to overcome this problem. This is referred to as the accuracy effect. To cater to our requirements, we apply bilinear interpolation with fractional-pel values. Given a non-integer position in the image, the intensities of its integer valued neighbours are adjusted such that on interpolation to the fractional position in between, the displaced intensity is obtained. The equation below indicates how the interpolation is performed on the left candidate image instead of Equation (1).

$$\delta = \lceil n - \alpha D_L(m,n) \rceil - [n - \alpha D_L(m,n)], \tag{3}$$

$$T_L(m, \lceil n - \alpha D_L(m,n) \rceil) = S_L(m,n), \tag{4}$$

$$T_L(m, \lfloor n - \alpha D_L(m,n) \rfloor) \;= \frac{S_L(m,n) - \delta T_L(m, \lceil n - \alpha D_L(m,n) \rceil)}{1-\delta}. \tag{5}$$

For the right candidate image, instead of Equation (2),

$$T_R(m, \lceil N-n+(1-\alpha)D_R(m,N-n) \rceil) = S_R(m,N-n), \tag{6}$$

$$\delta = \lceil N - n + (1-\alpha)D_R(m,N-n) \rceil - [N - n + (1-\alpha)D_R(m,N-n)], \tag{7}$$

$$T_R(m, \lfloor N - n + (1-\alpha)D_R(m,N-n) \rfloor) \;= \frac{S_R(m,N-n) - \delta T_R(m, \lceil N-n+(1-\alpha)D_R(m,N-n) \rceil)}{1-\delta}. \tag{8}$$

(a) (b)

Fig. 2. Results from Section 3.1. (a) Generated intermediate view (b) Disparity map of the generated view

(a) (b)

Fig. 3. (a) Final view generated from Section 3.2. (b) Final disparity map generated from Section 3.3.

Middlebury Dataset	Method in [8]		Using interpolation	
	PSNR(dB)	SSIM	PSNR(dB)	SSIM
Aloe	21.66	0.89	22.28	0.90
Art	25.06	0.89	26.37	0.92
Books	25.62	0.90	27.37	0.92
Cloth1	31.80	0.96	34.50	0.97
Dolls	27.76	0.91	29.70	0.94
Laundry	26.90	0.91	28.50	0.94
Moebius	27.63	0.91	29.18	0.93
Monopoly	24.83	0.91	26.11	0.92
Plastic	28.85	0.96	34.55	0.97
Rocks1	26.64	0.90	27.41	0.92

Table 1. Comparison of PSNR and SSIM values for initial view generation.

Using this interpolation technique gives higher PSNR and SSIM values as compared to the method in [8]. The results are shown in Table 1. The average PSNR over 27 datasets from the Middlebury Stereo Database is 28.81dB and the average SSIM is 0.93. An average PSNR gain of 2dB is obtained.

Intermediate disparity maps are also generated using the procedure described above. As in [8], the two candidate intermediate images and maps are then merged by placing pixels from both images into the new view and retaining pixels with greater disparity where pixels from both images occur at the same position. The image and its disparity map are then refined by applying a median filter along the edges where depth discontinuities occur. Fig. 2(a) and 2(b) show the results from this step. The double occlusion problem gives rise to holes in the generated view. These are filled in by applying techniques explained in the subsequent sections.

3.2. Filling holes in the disparity map

The holes in the disparity map occur both in foreground as well as in the background layers. Hence, it is necessary to apply the information from both these layers when filling in the holes. As suggested in [8], a histogram with B bins is formed from fixed size windows w (e.g., 31×31) surrounding the holes of the generated disparity map. In the event that the window is filled only with holes, the size of the window is increased to contain valid disparity values. The variance of the disparities, σ_w^2 for each window w is computed. A high variance indicates the presence of multiple disparity levels, while a low variance indicates a single depth layer. If multiple disparity values are seen, it is likely that an object in the foreground occludes the background and hence a lower disparity value is chosen. In case of lower variance, the common disparity value is placed in the hole. The cost function shown below is applied

$$l(b_i) = \beta \sigma_w^2 b_i + \frac{1}{c(b_i)},\ 1 \leq i \leq B, \tag{9}$$

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. Here we use a value of $\beta = 0.01$ and $B = 20$. When the first term in Equation (9) is larger compared to the second term, it is likely that the variance is high and hence a background value is chosen. When the second term is larger, a common disparity level dominates and hence this is chosen to be placed in the hole. A binary map B_{mn} is also created to indicate whether a given pixel position (m, n) is in the background ($B_{mn} = 1$) or else has a common disparity as its surrounding pixels ($B_{mn} = 0$). This is used in the next step for mapping pixels to the foreground or background layers.

Sudden changes at the edges of objects and depth discontinuities in the image are smoothed out by applying a 5×5 median filter (along the rows and columns) to each hole. This helps to improve the process of filling in the colours in the generated view in the subsequent step. Fig. 3(b) shows the disparity map generated from this step.

3.3. Filling holes in the generated view

Let x_{mn} indicate a missing pixel in the generated image. The binary map created in the previous step gives an indication B_{mn} about whether the pixel belongs in the foreground or in the background of the image. The values from the disparity maps of the original stereo views (D_L, D_R) are compared. If B_{mn} indicates a background value, the colour value from the

(a) (b)

Fig. 4. (a) Ground truth view (b) Absolute error image

Middlebury Dataset	Hole filling in [8]		Method proposed	
	PSNR(dB)	SSIM	PSNR(dB)	SSIM
Aloe	27.42	0.89	28.06	0.88
Art	31.53	0.94	30.22	0.94
Books	29.53	0.92	28.74	0.92
Cloth1	33.61	0.94	33.64	0.94
Dolls	32.73	0.94	30.90	0.94
Laundry	30.51	0.94	31.32	0.94
Moebius	33.17	0.93	32.76	0.93
Monopoly	29.89	0.94	28.77	0.93
Plastic	37.96	0.98	37.95	0.98
Rocks1	29.20	0.91	27.42	0.90

Table 2. Comparison of PSNR and SSIM values for hole filling in the generated view.

original stereo view with lower disparity at position (m, n) is placed in x_{mn}.

$$x_{mn} = \begin{cases} B_{mn}S_{L_{mn}} + (1-B_{mn})S_{R_{mn}} & ; D_{L_{mn}} < D_{R_{mn}} \\ B_{mn}S_{R_{mn}} + (1-B_{mn})S_{L_{mn}} & ; D_{L_{mn}} \geq D_{R_{mn}} \end{cases} \quad (10)$$

Else if $B_{mn} = 0$, then the colour with higher disparity value is given to x_{mn}. In this way, all the holes in the generated view are filled. Fig. 3(a) shows the result of this step. Table 2 compares the results of using the method for filling holes in the generated view from [8] combined with our interpolation technique, with the results of using our proposed method. The PSNR and SSIM values indicated here are averaged over views generated at $\alpha = 0.25$, $\alpha = 0.5$ and $\alpha = 0.75$. As compared to the hole filling technique described in [8], our method is simpler in terms of implementation and more efficient with lesser steps of execution.

4. RESULTS

The proposed algorithm was tested on 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 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 3.

Middlebury Dataset	PSNR(dB)			SSIM		
	View 2	View 3	View 4	View 2	View 3	View 4
Aloe	28.14	27.70	28.31	0.88	0.88	0.89
Art	30.65	29.91	30.05	0.94	0.94	0.94
Books	29.15	28.09	28.91	0.92	0.92	0.93
Cloth1	33.29	33.37	34.19	0.93	0.94	0.95
Dolls	30.96	30.55	31.16	0.94	0.94	0.94
Laundry	30.98	31.59	31.37	0.94	0.95	0.94
Moebius	32.69	32.72	32.87	0.93	0.93	0.93
Monopoly	28.93	27.92	29.35	0.94	0.93	0.94
Plastic	36.95	37.99	38.73	0.98	0.98	0.98
Rocks1	27.84	26.76	27.58	0.90	0.90	0.91

Table 3. PSNR and SSIM values for views with $\alpha = 0.25$, $\alpha = 0.5$ and $\alpha = 0.75$.

Middlebury Dataset	Method in [8]		Method proposed	
	PSNR(dB)	SSIM	PSNR(dB)	SSIM
Aloe	28.81	0.92	28.06	0.88
Art	31.67	0.95	30.22	0.94
Books	30.10	0.93	28.74	0.92
Cloth1	35.04	0.96	33.66	0.94
Dolls	31.61	0.95	30.90	0.94
Laundry	31.66	0.95	31.32	0.94
Moebius	33.42	0.95	32.76	0.93
Monopoly	29.80	0.95	28.77	0.93
Plastic	37.91	0.98	37.95	0.98
Rocks1	27.98	0.91	27.42	0.90

Table 4. Comparison of average PSNR and SSIM values with method in [8].

The results of the algorithm in [8] tested on the Middlebury datasets are shown in `http://videoprocessing.ucsd.edu/~ankitkj/research/viewsynthesis`. The average PSNR and SSIM values across the views generated are shown in Table 4 and are here compared with the average values obtained using the method from [8]. The average PSNR obtained across all the datasets was 31.84 dB and the average SSIM was 0.93. Fig. 4(a) shows the ground truth image and 4(b) shows the absolute error image for $\alpha = 0.5$.

The code was implemented entirely in Matlab and was run on an Intel P8400 on a single core with 4 GB of RAM. The image and disparity map resolutions used were $1110 \times 1240 - 1490$ pixels. The run time for the proposed method to generate a view and its disparity map is 50 seconds. A faster processor, such as an Intel i7 processor used in [8], with code translated into a compiled language will provide higher run times.

5. CONCLUSION

The paper presents a simple and efficient algorithm to generate content for multiview autostereoscopic displays. View synthesis using only a stereo pair introduces great possibilities in terms of reduction of the data transmitted to multiview autostereoscopic displays. A simple technique is presented here to create new views and this provides high PSNR and SSIM values which leads to better viewing quality.

Looking into the future, the next step is to extend this technique into non-rectified image data sets and produce views with the same quality. This in turn enables us to look into the possibility of generation of entire video streams from a new viewpoint.

6. REFERENCES

[1] H. Urey, K.V. Chellappan, E. Erden, and P. Surman, "State of the Art in Stereoscopic and Autostereoscopic Displays," *Proceedings of the IEEE*, vol. 99, no. 4, pp. 540–555, April 2011.

[2] D. Scharstein and C. Pal, "Learning Conditional Random Fields for Stereo," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Minneapolis, MN, USA, June 2007, pp. 1–8.

[3] H. Saito, S. Baba, M. Kimura, S. Vedula, and T. Kanade, "Appearance-based Virtual View Generation of Temporally-varying Events from Multi-camera Images in the 3D Room," in *Proceedings of the Second International Conference on 3-D Digital Imaging and Modeling*, Ottawa, Ont., Canada, 1999, pp. 516–525.

[4] C. Lipski, F. Klose, K. Ruhl, and M. Magnor, "The Virtual Video Camera: Simplified 3DTV Acquisition and Processing," in *3DTV Conference: The True Vision - Capture, Transmission and Display of 3D Video (3DTV-CON)*, Antalya, Turkey, May 2011, pp. 1–4.

[5] N.A. Manap and J.J. Soraghan, "Novel View Synthesis based on Depth Map Layers Representation," in *3DTV Conference: The True Vision - Capture, Transmission and Display of 3D Video (3DTV-CON)*, Antalya, Turkey, May 2011, pp. 1–4.

[6] A. Smolic, K. Muller, K. Dix, P. Merkle, P. Kauff, and T. Wiegand, "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, Oct. 2008, pp. 2448–2451.

[7] L. He, M. Bleyer, and M. Gelautz, "Object Removal by Depth-guided Inpainting," in *Proceedings of the OAGM/AAPR Workshop*, Graz, Austria, May 26-27 2011.

[8] A.K. Jain, L.C. Tran, R. Khoshabeh, and T.Q. Nguyen, "Efficient stereo-to-multiview synthesis," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Prague, Czech Republic, May 2011, pp. 889–892.