DEPTH IMAGE BASED RENDERING WITH ADVANCED TEXTURE SYNTHESIS

P. Ndjiki-Nya, M. Köppel, D. Doshkov, H. Lakshman, P. Merkle, K. Müller, and T. Wiegand

Fraunhofer Institut for Telecommunications, Heinrich-Hertz-Institut
Einsteinufer 37, 10587 Berlin, Germany
Email: {patrick.ndjiki-nya/martin.koeppel/dimitar.doshkov/haricharan.lakshman/
philipp.merkle/karsten.mueller/thomas.wiegand}@hhi.fraunhofer.de

ABSTRACT

In free viewpoint television or 3D video, depth image based rendering (DIBR) is used to generate virtual views based on a textured image and its associated depth information. In doing so, image regions which are occluded in the original view may become visible in the virtual image. One of the main challenges in DIBR is to extrapolate known textures into the disoccluded area without inserting subjective annoyance. In this paper, a new hole filling approach for DIBR using texture synthesis is presented. Initially, the depth map in the virtual view is filled at disoccluded locations. Then, in the textured image, holes of limited spatial extent are closed by solving Laplace equations. Larger disoccluded regions are initialized via median filtering and subsequently refined by patch-based texture synthesis. Experimental results show that the proposed approach provides improved rendering results in comparison to the latest MPEG view synthesis reference software (VSRS) version 3.6 [1].

Keywords— View synthesis, Texture synthesis, 3D video, Depth image based rendering.

1. INTRODUCTION

The interest in 3D video and free viewpoint television is constantly increasing and has led to improvements in all stages of the processing chain. Auto-stereoscopic displays provide a 3D impression to an observer without the need to wear additional glasses. Such a display shows a number of slightly different views (e.g. 9) at the same time, of which an observer sees only two in the right viewing positions. To simultaneously deliver so many videos, extremely large bandwidth is required. Additionally, there is a clear trend in the industry to use conventional stereo, thus providing two video streams with a camera distance optimized for cinemas. Hence, a need arises to render "virtual" views to support future displays at data rates almost similar to mono video [2].

Depth image based rendering (DIBR) can be used to render new views from a textured image and its associated depth values to a slightly different virtual view. One main problem that arises in this context is that the regions occluded by foreground (FG) objects in the original view may become visible in the "virtual" view. There are two basic options to handle this obstacle. The missing image regions may be replaced by meaningful color information [3] or the depth map may be preprocessed in a way that no disocclusions appear in the "virtual" image [4]. The disadvantage of existing approaches is that high quality rendering can be done just for small shifts of the camera position. There are two basic options to handle this obstacle. The missing image regions may be replaced by meaningful color information [3] or the depth map may be preprocessed in a way that no disocclusions appear in the "virtual" image [4]. The disadvantage of existing approaches is that high quality rendering can be done just for small shifts of the camera position.

Texture synthesis is an appropriate technique to fill unknown image locations with known information either from the same [5] or from other images in a database [6]. Texture synthesis operates in parametric [7] or non-parametric [5], [8] modes. While parametric methods are faster, non-parametric methods result in better visual quality [9].

In this paper, a new approach to handle disocclusions in 3D video is presented. The method is based on non-parametric texture synthesis such that "virtual" views with a large baseline can be rendered. A robust initialization gives an estimate of the unknown image regions that is further refined in a synthesis stage.

The remainder of this paper is organized as follows. In section 2, the overall algorithm is presented. In sections 3 to 5, depth map (DM) filling, image initialization and texture synthesis are presented in detail. In section 6, the simulation setup and experimental results are presented. Finally, conclusions and future steps are given in section 7.

2. OVERALL ALGORITHM

The proposed algorithm is depicted in Fig. 1 as a block diagram. The method introduced in [10] is used to compute the depth maps (DM). To warp an original image and its associated DM into the new "virtual" position, the method presented in [1] is utilized. As the "virtual" camera moves along the epipolar lines, the objects are moved in opposite direction along the same lines and some image regions are uncovered (cf. Fig. 2(a)).

The aim of the overall view synthesis algorithm is to cover the disoccluded area (holes) which becomes visible in the virtual view and the DM in a visually plausible manner. In a first step, the disoccluded area in the DM is filled. Then, very small holes are initially filled by solving Laplacian equations. The remaining holes are initialized from the spatially adjacent original texture to give an estimate of the missing information and patch-based texture synthesis is used to refine the initialized area.

We would like to thank the Gwangju Institute of Science and Technology in Korea and the Electronic and Telecommunications Research Institute / MPEG Korea Forum for providing the "Newspaper" and "Lovebird1" sequence respectively.

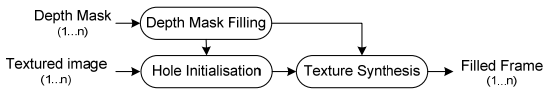

Fig. 1. Block diagram of the proposed approach. First, disocclusions in the DM are filled. Then the holes are initialized and refined with texture synthesis.

3. DEPTH MAP FILLING

The DM, denoted as D, is an 8 bit gray scale image. The closest point to the camera is associated with the value 255 and the most distant point is associated with the value 0 (cf. Fig. 2(a)-(c)). Depth filling is performed according to the reasonable assumption that the uncovered area belongs to BG and not to FG objects. The uncovered area in the DM is denoted as Ψ. Due to inaccuracies in depth estimation, FG objects may be warped into the disoccluded area. Therefore, blobs of up to γ pixels in Ψ are assigned to Ψ because they are potentially very noisy and may otherwise lead to noticeable inaccuracies in the post-processed DM (cf. Fig. 2(b)). The holes in the DM are filled line-wise along the epipolar lines from the BG. The filling direction for the example in Fig. 2(a) is marked with red arrows. One possibility is to copy the last valid BG depth value D_i line-wise into Ψ (cf. Fig. 2(b)). But relying on a single D_i value is not likely to be robust. For that reason, two centroids (c_{min} and c_{max}) representing the FG and BG of D_i's neighborhood are computed via k-means clustering ($k = 2$). The considered neighborhood is determined by a window of size $M \times N$ pixels, centered on the location D_i. The condition to choose the depth value to be filled is as follows (cf. Fig. 2 (c)):

$$D_j = \begin{cases} D_i \text{ if } D_i \leq c_{min} \\ c_{min} \text{ otherwise} \end{cases}, \quad j \in \Psi; i \in D/\Psi. \tag{1}$$

4. INITIAL FILLING OF TEXTURED PICTURES

In an initial filling step, small holes in the current picture are covered by using the Laplacian equation [11], which works well for the reconstruction of smooth regions. It can be assumed, that a small area of missing information fulfills this condition. This restoration method provides good visual results for holes less than γ pixels (e.g. 50 pixels) and is faster than patch-based texture synthesis. This area is considered as finally filled and will not be refined with texture synthesis.

In [12], it is shown that the performance of texture synthesis can be improved by using an initial estimate of pixel values in the unknown region. In this paper, an initialization scheme that is based on the statistical properties of known samples in the vicinity of the hole area is employed. Generally, the known samples constitute valid BG pixels but in some cases the depth information at the FG-BG transition are not reliable. Hence, the probability distribution of known BG pixel values in the spatial neighborhood of the hole area is observed to be skewed. To capture the BG value from the spatially adjacent samples, we utilize the Median estimator, which is the standard measure of (end value) location used in case of skewed distributions.

A window of samples centered on the pixel to be filled is considered. For each unknown pixel, a measure N_{BG} is set equal to the number of known pixels that are classified as BG in the current window. The unknown pixels are visited in decreasing order of N_{BG}. A 2D median filter operates on the BG pixels in the current window and the filtered output is used to initialize the unknown pixel. The filtering operation can be viewed as the process of extracting a valid BG value from the spatially neighboring samples.

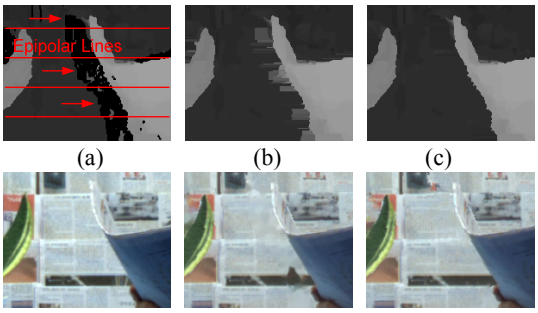

Fig. 2. Results for frame 1 of the "Newspaper" sequence for the proposed depth map (DM) and texture filling approach. (a) DM with disoccluded area marked black (filling direction given by red arrows). (b) Line-wise filling of DM without blob removal. (c) Result of proposed DM filling approach. (d) Original reference image. (e) Result by MPEG VSRS. (f) Results by the proposed approach.

This serves as a coarse estimate for the texture synthesis stage that can bring back the details in the unknown region. Using the described initialization scheme, the sensitivity of the texture synthesis stage to outliers in the transition region is significantly reduced.

5. TEXTURE SYNTHESIS OF TEXTURED PICTURES

Patch based texture synthesis is a process in which small patches from a known area are used to fill an unknown area. The patch to be copied overlaps with original/synthesized samples such that there is a smooth fit. The algorithm proposed in [5] is utilized to determine the filling order and is enhanced in two ways. Firstly, already estimated samples through initialization are considered in further processing steps. The gradient is computed for all the samples in the current patch, thus leading to a better isophote direction (Please refer to [5] for the original algorithm). Secondly, the filling order is steered such that the synthesis starts from the BG area towards the FG objects. To this end, locations in Ψ are assigned filling priorities correspondingly. In the following, $\boldsymbol{c}$ denotes the patch at the current location to be filled, whose center is denoted as $\boldsymbol{c}_{center}$. An area of $5M \times 5N$ centered at $\boldsymbol{c}_{center}$ is defined to be the source area $\boldsymbol{s}$. The filling algorithm now searches for a patch $\boldsymbol{x}$ in $\boldsymbol{s}$ that is similar to $\boldsymbol{c}$. Only the luminance samples are considered in the matching routine. Using the DM that is already filled, the depth value of $\boldsymbol{c}_{center}$ is always known. All pixel positions in $\boldsymbol{s}$ with depth values higher than $D_{\boldsymbol{c}_{center}} + \beta$ are excluded from search. In such a manner, patches will not be taken from area with depth values much higher than the current region to be filled, in other words foreground objects. The source area is sub-sampled by a factor of 2 to accelerate the matching operation. The remaining source positions are utilized as center positions for $\boldsymbol{x}$. The best continuation patch out of all candidate patches is found by minimizing the following cost function:

$$E = \sum_{i=1}^{K} \|\boldsymbol{x}_i - \boldsymbol{c}_i\|^2 + \omega_\Omega \sum_{j=1}^{K_\Omega} \left\|\boldsymbol{x}_j - \boldsymbol{c}_j\right\|^2 + \omega_\nabla \sum_{i=1}^{K} \|\nabla \boldsymbol{x}_i - \nabla \boldsymbol{c}_i\|^2 + \omega_\nabla \omega_\Omega \sum_{j=1}^{K_\Omega} \left\|\nabla \boldsymbol{x}_j - \nabla \boldsymbol{c}_j\right\|^2 \tag{2}$$

where K is the number of original and K_Ω is the number of initialized pixels in $\boldsymbol{c}$. $\nabla \boldsymbol{c}$ is the sample-wise gradient of $\mathbf{c}$ and $\nabla \boldsymbol{x}$ is the sample-wise gradient of $\boldsymbol{x}$. ω_Ω is the weighting factor for the initialized values in Ω and ω_∇ is the weighting factor for gradient component in the matching process. In the last term the weight is given

by $\omega_\Omega \omega_\nabla$ because here, the distance between $\boldsymbol{x}$ and $\boldsymbol{c}$ is determined both in sample and gradient domain. An efficient post-processing method is applied to ensure smooth transitions between adjacent patches [8]. This post-processing is adapted to the framework in such a manner that FG objects are not considered as boundary pixels.

6. SIMULATION SETUP AND EXPERIMENTAL RESULTS

To evaluate the proposed approach, we used three 3D-video sequences "Book arrival", "Lovebird1" and "Newspaper", having a resolution of 1024 x 768 pixels. For these sequences several rectified videos with slightly different camera perspectives are available. The baseline is approximately 65 mm between two adjacent cameras. For every sequence, we considered 2 original but not necessarily adjacent cameras (cf. Table 1 "Camera" column). The following view synthesis operations were conducted: warping an original view (right and left) towards an adjacent view; baseline extension where the virtual camera position is 2 cameras away from the original camera location, giving a baseline of approximately 130 mm. For all experiments we set $M = N = 32$ pixels, $\gamma = 50$ pixels, $\omega_\Omega = \omega_\nabla = 0.6$ and $\beta = 15$. Objective results are depicted in Table 1 by the mean PSNR and SSIM over all pictures of a sequence. PSNR is computed locally only for the defective area in the image while SSIM is determined for the entire image, because application of SSIM to arbitrarily shaped regions is not straightforward. Note that camera 7 of "Lovebird1" and camera 5 of "Newspaper" were not available so we had no reference for objective evaluation of the middle view case. For the sequence "Book arrival" the reported approach gives better SSIM and PSNR results than MPEG VSRS. For the "Lovebird1" sequence we obtain the best results in terms of SSIM. However, the PSNR value of MPEG VSRS is better for the case "camera 6 to 8" because although VSRS yields obvious artifacts, the color information in the synthesized area seams to correlate with the true information and is strongly blurred, while our result is sharper but noisy. For the sequence "Newspaper" VSRS gives the better results for the synthesis of camera 6 from camera 4, because our modules rely on the DM. However, here, the DM is particularly unreliable yielding visual and objective losses. Nevertheless, objective and visual gains can be reached for the case "camera 6 to 4" as shown in Table 1 and in Fig. 2(d)-(f) (electronic magnification may be required) respectively. In Fig. 2(d), the original reference image is shown. Fig. 2(e) shows the result with VSRS and Fig. 2(f) the results with the proposed software. Fig. 3 exemplarily shows objective results in PSNR and SSIM for one warping direction with large baseline out of all sequences. In Fig. 4 and Fig. 5 visual results for the sequences "book arrival" and "lovebird" are shown. In (a), the original reference picture is shown. It is frame 1 of the sequence "Book arrival" and frame 116 for "Lovebird1". In (b), the warped images are shown (large baseline extension for "Book arrival" camera 8 to 10, for "Lovebird1" camera 8 to 6). The disoccluded areas are marked black. The rendering results by MPEG VSRS are shown in (c), while the rendering results by the proposed approach are shown in (d). Fig. 4 (e), (f) and Fig. 5 (e), (f) are enlargements of the red bounding boxes shown in (c) and (d), where the results for the proposed algorithm are shown on the right and the MPEG VSRS results are depicted on the left. In Fig. 4 (e), (f) and Fig. 5 (e), (f) it can be seen that our algorithm correctly fills the disoccluded areas, as no foreground data is used. This also leads to better object boundaries. As can be seen in Fig. 4 (f) on the poster in the background, details are well preserved by our method. Also in Fig. 5 (f), the staircase is reconstructed in a visually plausible manner by the proposed approach.

Table 1. PSNR and SSIM results by the proposed framework and the view synthesis reference software.

Seq.	Camera	PSNR (dB)		SSIM	
		Prop.	MPEG	Prop.	MPEG
Book.	8 to 9	37.07	36.06	0.9833	0.9828
Book.	10 to 9	36.22	35.15	09838	0.981
Book.	8 to 10	31.08	30.25	0.954	0.9525
Book.	10 to 8	30.77	30.29	0.9555	0.9524
Love.	6 to 8	39.24	42.13	0.9286	0.9284
Love.	8 to 6	39.26	38.54	0.9433	0.9425
New.	4 to 6	25.07	25.43	0.8969	0.8974
New.	6 to 4	31.25	30.37	0.9149	0.9131

7. CONCLUSIONS AND FUTURE WORK

In this paper a new hole filling algorithm for DIBR is presented. The approach works for large baselines and the rendering results are visually consistent. A robust initialization is used to obtain an estimate of the disoccluded area. Subsequently, a refinement step based on patch-based texture synthesis is applied. Overall, the proposed algorithm gives both subjective and objective gains compared to the latest MPEG VSRS. However, all modules of the approach depend on the depth map which can be especially unreliable at background-foreground transitions. Hence, wrong depth estimates may lead to significant degradation of the rendering results. In future work this dependency will be relaxed. Furthermore, alternative state-of-the-art synthesis methods will be examined in order to assess the potential gains that can be expected from texture synthesis algorithms in the DBIR context.

8. REFERENCES

[1] M.Tanimoto, T. Fujii, and K. Suzuki, "View Synthesis Algorithm in View Synthesis Reference Software 2.0 (VSRS2.0)", ISO/IEC JTC1/SC29/WG11 M16090, Lausanne, Switzerland, February 2008.

[2] A. Smolic, K. Müller, and A. Vetro, "Development of a New MPEG Standard for Advanced 3D Video Applications", *In Proc. of IEEE Int. Symp. on Image Signal Processing and Analysis,* Salzburg, Austria, September 2009.

[3] C.-M. Cheng, S.-J. Lin, S.-H- Lai, and J.-C. Yang, "Improved Novel View Synthesis from Depth Image with Large Baseline", *In Proc. of Int. Conf. on Pattern Recognition*, Tampa, USA, December 2008.

[4] C. Fehn, "Depth Image Based Rendering (DIBR), compression and transmission for a new approach on 3D-TV", *In SPIE Stereoscopic Displays and Virtual Reality Systems XI*, pp. 93-104, January 2004.

[5] A. Criminisi, P. Perez, and K. Toyama, "Region Filling and Object Removal by Exemplar-based Inpainting", *In IEEE Trans. on Image Proc.*, vol. 13, no. 9, pp. 1200-1212, January 2004.

[6] J. Hayes, A. Efros, "Scene Completion Using Millions of Photographs", *In Proc. ACM SIGGRAPH,* San Diego, USA, August 2007.

[7] G. Doretto, A. Chiuso, Y. N. Wu, and S. Soatto, "Dynamic Textures", *In Int. Journal of Com. Vision*, pp. 91-109, February 2004.

[8] P. Ndjiki-Nya, M. Köppel, D. Doshkov, and T. Wiegand, "Automatic Structure-Aware Inpainting for Complex Image Content", In Proc. of Int. Sym. on Visual Computing, Las Vegas, USA, December 2009.

[9] L.-Y. Wei, S. Lefebvre, V. Kwatra, and G. Turk, "State of the Art in Example-based Texture Synthesis" *EUROGRAPHICS 2009, State of the Art Report, EG-Star*, Munich, Germany, 2009.

[10] M.Tanimoto, T. Fujii, and K. Suzuki, "Depth Estimation Reference Software (DERS) 5.0", ISO/IEC JTC1/SC29/WG11 M16923, Lausanne, Switzerland, October 2009.

[11] P. Pérez, M. Gangnet, and A. Blake, "Poisson Image Editing", *In Proc. ACM SIGGRAPH*, San Diego, USA, July 2003.

[12] H. Lakshman, M. Köppel, P. Ndjiki-Nya, and T. Wiegand, "Image Recovery Using Sparse Reconstruction Based Texture Refinement", *In Proc of IEEE Int. Conf. on Acustic Speech and Signal Proc.*, Dallas, USA, March 2010.

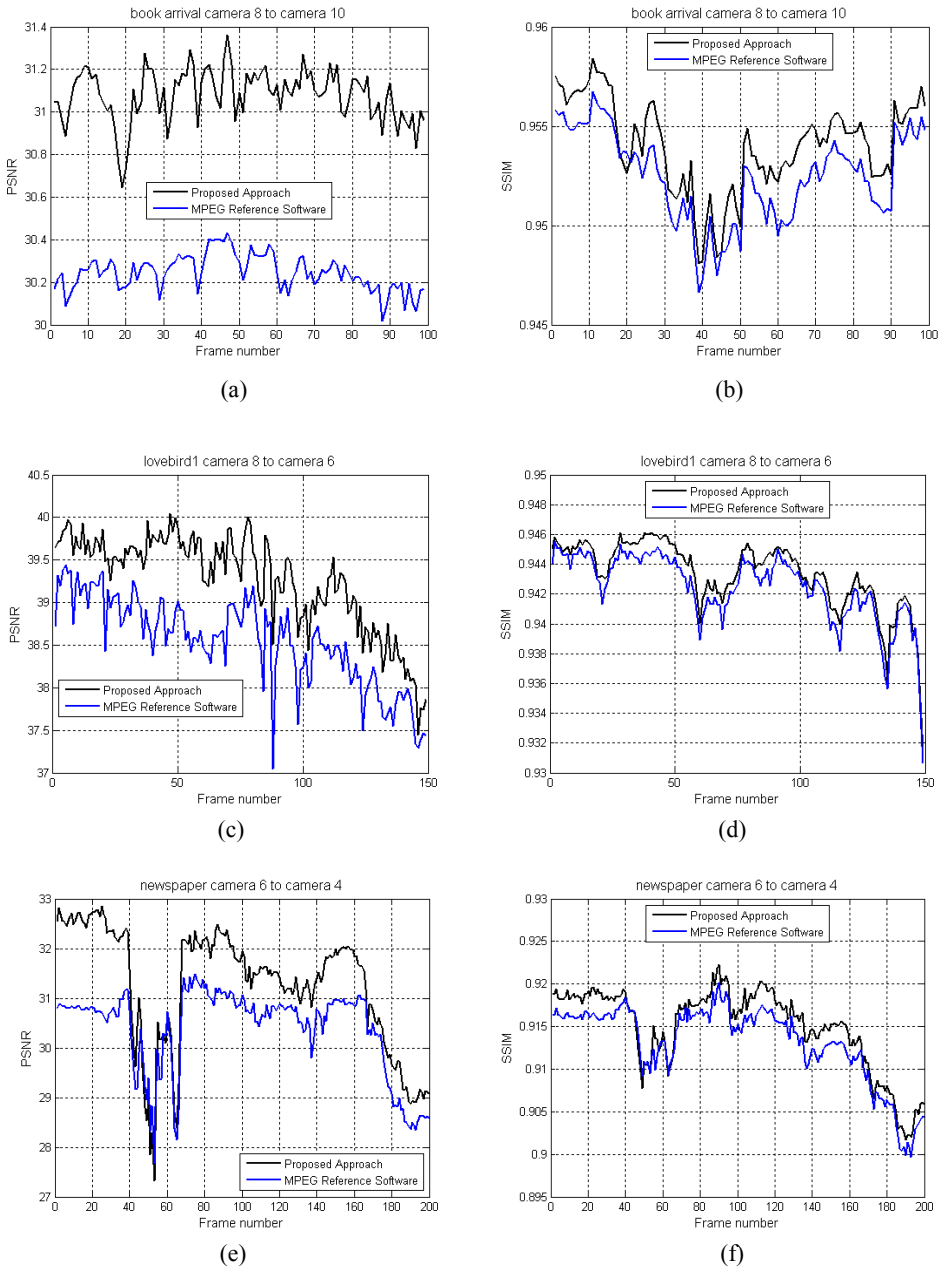

Fig. 3. Objective results for the "Book arrival", "Lovebird1" and "Newspaper" sequences. (a), (c), (e) PSNR for all pictures of the sequence measured locally in the defected area. (b), (d), (f) SSIM for all pictures of the sequence measured for the entire image.

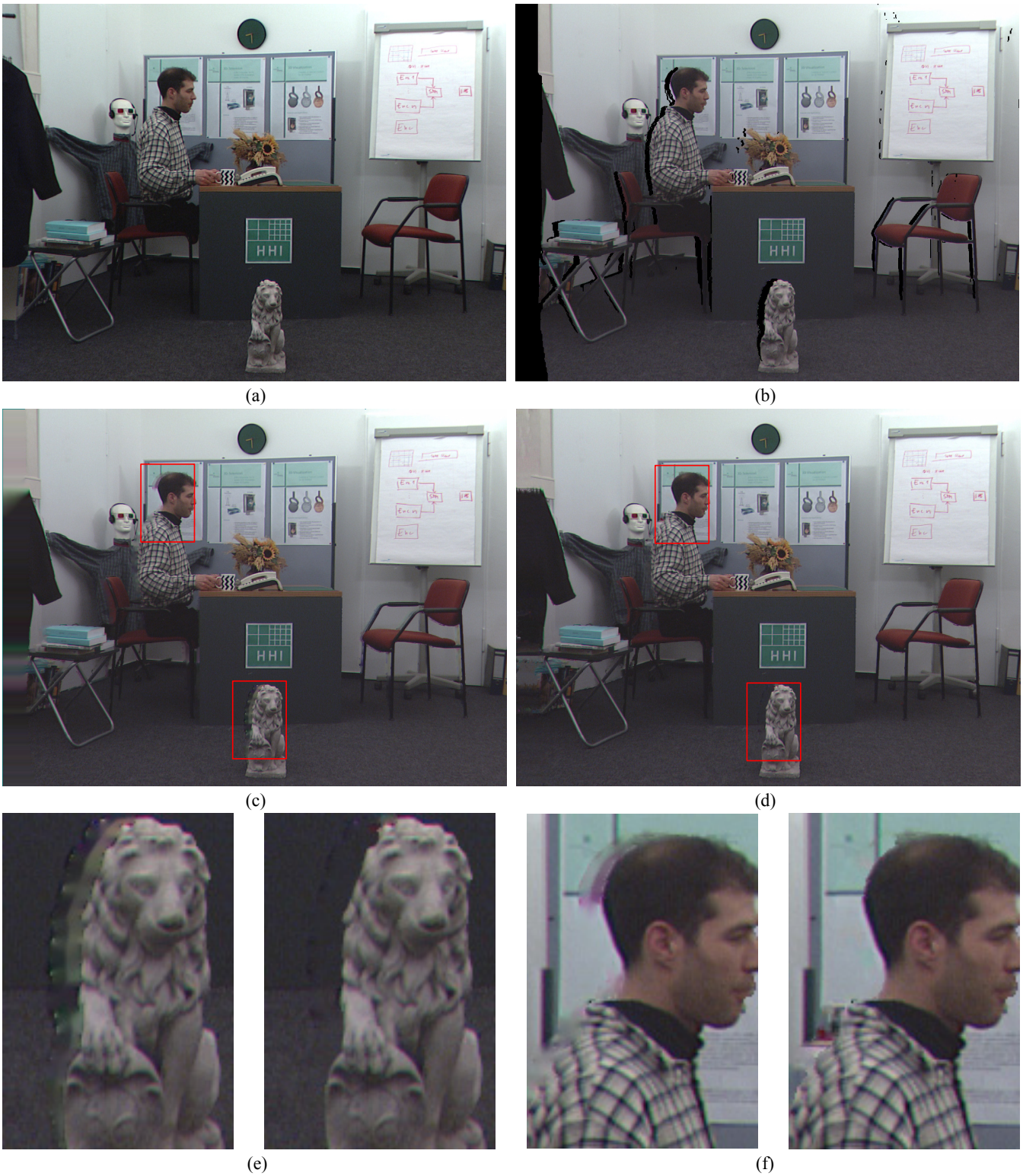

Fig. 4. DIBR results for the sequence "Book arrival". (a) Original reference image. (b) Warped image with uncovered area marked black. (c) Result of picture 1 by MPEG VSRS. (d) Result of the proposed approach for the same picture. (e) and (f) Magnified results. Left, MPEG VSRS. Right, the proposed approach.

Fig. 5. DIBR results for the sequence “Lovebird1”. (a) Original reference image. (b) Warped image with uncovered area marked black. (c) Result of picture 116 by MPEG VSRS. (d) Result of the proposed approach for the same picture. (e) and (f) Magnified results. Left, MPEG VSRS. Right, the proposed approach.