Consistent Stereo Image Editing

Tao Yan[1,2], Shengfeng He[2], Rynson W.H. Lau[2] and Yun Xu[1]
[1]University of Science and Technology of China, China
[2]City University of Hong Kong, Hong Kong
yantao@mail.ustc.edu.cn, shengfeng_he@yahoo.com,
rynson.lau@cityu.edu.hk, xuyun@ustc.edu.cn

ABSTRACT

Stereo images and videos are very popular in recent years, and techniques for processing this media are attracting a lot of attention. In this paper, we extend the shift-map method for stereo image editing. Our method simultaneously processes the left and right images on pixel level using a global optimization algorithm. It enforces photo consistence between the two images and preserves 3D scene structures. It also addresses the occlusion and disocclusion problem, which may enable many stereo image editing functions, such as depth mapping, object depth adjustment and non-homogeneous image resizing. Our experiments show that the proposed method produces high quality results in various editing functions.

Categories and Subject Descriptors

I.4 [**IMAGE PROCESSING AND COMPUTER VISION**]: Enhancement

Keywords

Stereo Images Editing, Depth Mapping, Multi-label Optimization

1. INTRODUCTION

With the popularity of stereo (or 3D) videos, techniques for producing and editing of stereo media are attracting a lot of research attention in recent years. In general, a stereo image contains two regular 2D images, which are captured from the same scene at the same time but from slightly different locations. A naive solution for editing a stereo image is to separately process the left and right images and then recombine them to form the output stereo image. This will, however, introduce inconsistences between the left and right images. To address this problem, we may apply a global optimization algorithm to simultaneously process the left and right images to ensure consistence, e.g., [6] [11].

Most existing stereo editing works [3] [4] [6] [8] [7] [9] [11] are based on mesh warping, which treat the whole left or right image as a continuous plane by fitting a quad or triangle mesh to it and warp the shape of the mesh in a content-aware manner. Although this kind of methods are usually robust, they suffer from two major limitations. First, they cannot handle object boundaries well. Second, they cannot support editing functions that introduce new occlusion or disocclusion. As a result, some image editing functions such as depth mapping, object depth adjustment and non-homogeneous image resizing cannot be easily supported by mesh warping based methods. To address these limitations, we propose a novel framework for consistent stereo image editing in this paper. It is based on the shift-map method [10] and the multi-label optimization method [2]. The basic-idea of shift-map is that a pixel (x,y) in the output image can be obtained from a pixel $(x+t_x, y+t_y)$ of the input image. Hence, the output image can be computed as:

$$I'(x,y) = I(x+t_x, y+t_y) \tag{1}$$

where I is the input image. I' is the output image. A main objective of shift-map is to address the limitations of the mesh warping methods in regular 2D image editing.

The main contribution of this work is to extend the shift-map method for stereo image editing. The proposed method processes both the left and right images simultaneously to preserve photo consistence and minimize distortion in image editing. It is also able to handle occlusion and disocclusion, which are introduced by the editing operations, through a revised multi-label optimization algorithm. This allows the proposed method to support many different stereo editing functions, including depth mapping, object depth adjustment and non-homogeneous stereo image resizing.

The rest of this paper is organized as follows. Section 2 briefly summarizes related works on stereo image processing. Section 3 presents our consistent stereo image editing method. Section 4 discusses some experimental results and image editing functions.

2. RELATED WORK

Techniques for stereo image/video editing can be roughly classified into continuous methods and discrete methods. They are summarized as follows.

Continuous Methods: This type of methods is mainly based on mesh warping. Lang et al. [6] propose a disparity mapping method with four simple operators for adjusting the disparity range of stereo videos. However, their method does not consider the relative depths among objects in the stereo image and the relationship between depth perception and viewing parameters, such as display distance and size. They also do not consider 3D scene structure preservation. To overcome the display adaptation problem, Yan et al. [11] propose a depth mapping method that applies 3D image analysis techniques to remap the depth range of stereo videos according to the viewing parameters. Their method attempts to preserve 3D scene structures and enforce depth and temporal coherences through a global optimization algorithm.

There is some work on non-homogeneous stereo image resizing. Chang et al. [3] propose a mesh warping based method for resizing stereo images according to the change in viewing parame-

MM'13, October 21 - 25 2013, Barcelona, Spain
Copyright 2013 ACM 978-1-4503-2404-5/13/10 ...$15.00.
http://dx.doi.org/10.1145/2502081.2502177

ters. However, their method may introduce distortion to the shape of prominent objects. Lee et al. [7] fit a separate mesh to each depth layer of a stereo image, but they do not handle the disocclusion problem. In addition, they also rely on the user to segment the stereo image into different depth layers. There is also some work to develop specific stereo image editing techniques. Luo et al. [8] present a method for seamless stereo image cloning, which involves shape adjustment of the target object according to the depth and the color of the background stereo image. Tong et al. [4] propose a depth-consistent stereo image composition method, which allows interactive blending of a 2D image object on a stereo image. Niu et al. [9] extend the traditional 2D image warping technique for stereo images. Their objective is to preserve prominent objects and 3D scene structure.

Although warping based methods can produce visually pleasing results for many image processing functions, the main drawback is that they cannot handle object boundaries (or discontinuity) well. As a result, they cannot model scene occlusion and disocclusion.

Discrete Methods: Basha et al. [1] extend the seam carving method to non-homogeneous stereo image resizing. Their method simultaneously carves a pair of piece-wise continuous seams in both images, while minimizing distortion in appearance and depth. However, its performance is affected by the noise in the disparity map. In addition, as seam carving is based on dynamic programming, the result is not globally optimal. On the other hand, our method is a global optimization method and it is able to work with moderate quality disparity maps. Our method can also be applied to many stereo editing functions, in addition to non-homogeneous stereo image resizing.

The shift-map method [10] characterizes geometric rearrangement of a 2D image by a shift-map. It is a general tool for 2D image editing, such as resizing, composition and object removal. As it is not designed for handling stereo images, it does not consider consistence between the left and right images when applied to stereo images. It also cannot automatically handle occlusion and disocclusion introduced by image editing. In addition, it does not consider 3D structure correctness and preservation as it is just for 2D image editing. Our method extends the shift-map method to handle stereo images and addresses the above limitations.

3. OUR METHOD

We first denote the input stereo image pair as I_L (left image) and I_R (right image), their corresponding disparity maps as D_{LR} and D_{RL}, the output image pair as I'_L and I'_R, and their shift-maps as M_L and M_R, respectively. We formulate the stereo image editing problem as an energy minimization problem as follows:

$$E = E_{dc} + \omega_{sc} E_{sc} + \omega_{ph} E_{ph} + \omega_{pl} E_{pl} \tag{2}$$

where E_{dc} is the data cost for pixel rearrangement. E_{sc} is the smooth cost between neighboring pixels within the left or right image separately. E_{ph} is used to enforce photo consistence between the left and right images. E_{pl} is used to preserve 3D planes in the stereo image. $\omega_{sc} = 30$, $\omega_{ph} = 1$ and $\omega_{pl} = 10^4$ are weights used in our implementation. E_{dc} and E_{sc} are also used in [10] and are designed for 2D image editing. We modify them for stereo image editing. E_{ph} and E_{pl} are two new energy terms that we introduce to enforce photo consistence and feature preservation, respectively, in stereo image editing.

We apply the multi-label optimization method [2] to solve the above energy function. Each pixel in the output image is assigned a label, which is a vector indicating its location in the input image. The output stereo image can then be obtained according to the computed shift maps and the input stereo image.

3.1 Disparity Map Computation

In stereo image editing, we often need to determine the relationship between the left and right images. We compute the left to right D_{LR} and right to left D_{RL} disparity maps for the input stereo image pair using the method suggested by [2]. Since some disparity values, such as those in the occluded regions or near to the image boundary, may be incorrect, we perform a left-right consistence check on the computed disparity maps and delete those values that are inconsistent. We would like to emphasize here that the disparity maps needed in our method do not need to be very dense.

3.2 Data Cost

A lot of image editing functions involve pixel rearrangement. Energy term E_{dc} is used to model this data cost. To extend it to stereo images, we formulate this energy term as:

$$\begin{aligned} E_{dc} = \sum_{p \in I'_L \bigcup I'_R} & w_s(w_l min(\|t_p - g(p)\|^2, \tau_L) + min(\|I(p+t_p) - \\ & \hat{I}(p)\|^2, \tau_I) + min(\|D(p+t_p) - \hat{D}(p)\|^2, \tau_D)) \end{aligned} \tag{3}$$

where t_p is the label of pixel p in the output image. $g(p)$ is the target label of p, which is known in advance. w_S is the importance of pixel $p+t_p$ in the input image. The importance maps of a stereo image are computed by dividing the saliency of each pixel by its normalized disparity value. We then enhance the importance maps with user defined masks. $\hat{I}(p)$ and $\hat{D}(p)$ are the target color and disparity of p. In stereo image editing, it is possible that some pixels may move to the same location in the output image. Since only the nearest pixels should be visible, we use a depth buffer to determine the values of $\hat{I}(p)$ and $\hat{D}(p)$.

3.3 Smooth Cost

Energy term E_{sc} is used to enforce the smoothness of labels among neighboring pixels. Pritch et al. [10] only use color and image gradient information to compute the smooth cost of neighboring pixels. Since stereo images contain depth information, we extend the smooth cost to $\mathbf{V} = \langle I, \bigtriangledown I, D \rangle$, where I is the color information, $\bigtriangledown I$ is image gradient, D is the disparity value. We define our smooth penalty function as:

$$E_{sc} = \sum_{(p,q) \in N_L \bigcup N_R} min(\mathbf{w}_v^T (|\mathbf{V}_{p+t_p} - \mathbf{V}_{n'_q}| + |\mathbf{V}_{n'_p} - \mathbf{V}_{q+t_q}|), \tau_{sc}) \tag{4}$$

where N_L and N_R refer to neighboring pixel pairs in the left and right output images. Here, we only consider 4-connected pixel neighborhood. Consider two neighboring pixels, $p = (x_1, y_1)$ and $q = (x_2, y_2)$. t_p and t_q denote labels of p and q. $n'_p = (x_1 + t_{x_1} + 1, y_1 + t_{y_1,})$ and $n'_q = (x_2 + t_{x_2} - 1, y_2 + t_{y_2})$, if $x_1 < x_2$ and $y_1 = y_2$. $n'_p = (x_1 + t_{x_1}, y_1 + t_{y_1} + 1)$ and $n'_q = (x_2 + t_{x_2}, y_2 + t_{y_2} - 1)$, if $x_1 = x_2$ and $y_1 < y_2$.

We would like to point out here that the smooth cost defined in [10] is not suitable for us, as it does not satisfy the regularity constraint [5]: $E^{i,j}(0,0) + E^{i,j}(1,1) \leq E^{i,j}(0,1) + E^{i,j}(1,0)$. As such, the optimization step may not produce acceptable minimal energy values in stereo image editing. Obviously, our smooth cost is regular and metric [2] [5], and does not have this problem.

3.4 Photo Consistence

In a stereo image, each pair of matched pixels from the left and right images should have similar colors, and we use this matched pixel pair to estimate its depth. Energy term E_{ph} is used to ensure photo consistence between the left and right images through checking the color/depth consistence.

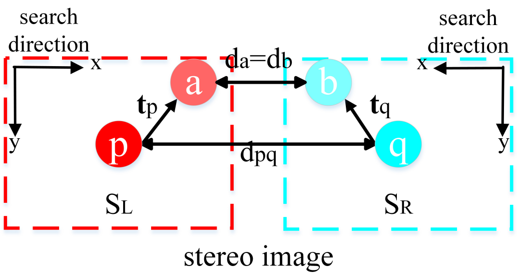

Figure 1: To find a candidate matched pixel pair in depth mapping. Pixel a is within search window S_L of p in the left image; pixel b is within search window S_R of q in the right image. If a and b are a corresponding pixel pair in the input images, $\mathbf{t}_p = \mathbf{t}_q$, and d_{pq} is within an acceptable disparity range, then p and q are considered as a candidate matched pixel pair in output images.

Our objective here is to produce a stereo output image that satisfies the photo consistence constraint. However, it is difficult to determine matched pixel pairs in the output stereo image before we obtain this output image. Here, we propose a strategy to dynamically determine matched pixel pairs for the output image in our global optimization step. We first select some candidate matched pixel pairs from the left image to the right image and some from the right image to the left image. We then put all these candidate matched pixel pairs into the global optimization process to determine the final matched pixel pairs.

As shown in Fig. 1, let p be a pixel in I'_L and q be a pixel in I'_R. We define the search windows for p and q as S_L and S_R. Search window S_L (or S_R) contains all pixels in the input image I_L (or I_R) that can be shifted to p (or q) in the output image I'_L (or I'_L). For different editing functions, the search directions (the order of labeling) of S_L and S_R may be different. For depth mapping, the search direction of S_L is from left to right and top to bottom, while that of S_R is from right to left and top to bottom. This is to satisfy the depth mapping constraint [11], and regularity and metric constraint.

We first consider how to determine the candidate matched pixel pairs from the left image to the right image. Let $a = I_L(x,y)$ be a pixel within S_L of p, where $x = x_p + t_{x_p}$ and $y = y_p + t_{y_p}$. Let $b = I_R(x + D_{LR}(x,y), y)$ be a pixel in S_R of q. If $D_{LR}(x,y) = D_{RL}(x + D_{LR}(x,y), y)$ and $|I_L(x,y) - I_R(x + D_{LR}(x,y), y)| < \phi$, then b is said to be the corresponding pixel of a. Now, if the distance $(2t_{x_p} + D_{LR}(x,y))$ between $p = (x_p, y_p)$ and $q = (x_p + 2t_{x_p} + D_{LR}(x,y), y_p)$ is within an acceptable disparity range, then we assume that p and q are a candidate matched pixel pair. (Here, an acceptable disparity range for pixels p and q in the output stereo image is defined as the range from $D_{LR}(a)$ to the user specified maximum disparity for a.) Finally, if there are no edges existed between p and q, we add an undirected edge between them.

We then consider how to determine the candidate matched pixel pairs from the right image to the left image. Let $b = I_R(x,y)$ be a pixel within S_R of q, where $x = x_q - t_{x_q}$ and $y = y_q + t_{y_q}$. Let $a = I_L(x - D_{RL}(x,y), y)$ be a pixel in S_L of p. Similarly, if $D_{RL}(x,y) = D_{LR}(x - D_{RL}(x,y), y)$ and $|I_R(x,y) - I_L(x - D_{RL}(x,y), y)| < \phi$, then a is said to be the corresponding pixel of b. Again, if the distance $(2t_{x_q} + D_{RL}(x,y))$ between $q = (x_q, y_q)$ and $p = (x_p - 2t_{x_q} - D_{RL}(x,y), y_q)$ is within an acceptable disparity range, then we assume that q and p are a candidate matched pixel pair. Finally, if there are no edges existed between p and q, we add an undirected edge between them. The penalty of undirected edge (p,q) is defined as:

$$\phi_{ph}(p,q,t_p,t_q) = \begin{cases} 0, & \text{if } t_p = t_q \\ K, & \text{else} \end{cases} \tag{5}$$

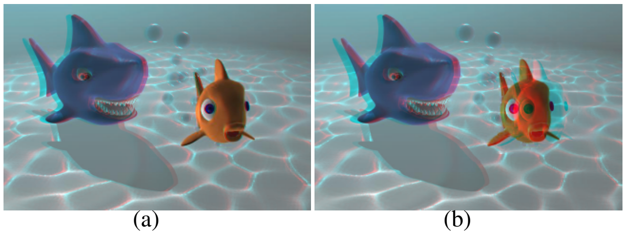
(a) (b)

Figure 2: Depth mapping: (a) original stereo image; (b) our result.

where, $\phi = 30$ and $K = 10^2$ in our implementation. Obviously, this penalty definition is metric [2]. Thus, E_{ph} is defined as:

$$E_{ph} = \sum_{(p,q)\in M_{LR} \bigcup M_{RL}} \phi_{ph}(p,q,t_p,t_q) \tag{6}$$

where M_{LR} and M_{RL} are candidate matched pixel pairs from the left image to the right and the right image to the left, respectively.

3.5 3D Feature Preservation

In stereo image editing, 3D image features should be carefully preserved. Since edge preservation is already considered in our smooth cost function. We mainly consider plane preservation.

We use P_l and P_r to represent the same plane projected in the left and right images. Pixels within P_l should satisfy constraint $d = a_l x + b_l y + c_l$, and pixels within P_r should satisfy constraint $d = a_r x + b_r y + c_r$, where d is the disparity of (x,y). a_l, b_l, c_l and a_r, b_r, c_r are parameters of the plane equations, which can be determined by [11].

In our current implementation, we manually define a mask on the input left image to obtain a plane mask. The plane preservation mask for right image is then computed automatically. We use the method presented in [11] to compute the target coordinates $(\hat{x}_{i,l}, \hat{y}_i)$ and $(\hat{x}_{i,r}, \hat{y}_i)$ of each matched pixel pair $(x_{i,l}, y_i)$ and $(x_{i,r}, y_i)$ on P_l and P_r. We define the plane preservation penalty as:

$$\begin{aligned} E_{pl} = \sum_{(P_l,P_r)} \sum_i \{ & min(|t_{i,x,l} - (\hat{x}_{i,l} - x_{i,l})|, 1) + min(|t_{i,x,r} - \\ & (\hat{x}_{i,r} - x_{i,r})|, 1) + min(|t_{i,y} - (\hat{y}_i - y_i)|, 1)\} \end{aligned} \tag{7}$$

4. RESULTS AND APPLICATIONS

We have implemented the proposed method in C and OpenCV. In our experiments, we used a $19in$ screen with a resolution of 1280×1024 and pixel density of $\beta = 0.293mm/pixel$. Other parameter setting include $w_l = 100$, $\tau_L = 10$, $\tau_I = 10^4$, $\tau_D = 2 \times 10^2$, $\mathbf{w}_v = \langle 1, 10, 1 \rangle$, and $\tau_{sc} = 10^4$. We ran the program (unoptimized) on a PC with an i5 3.1GHz CPU and 8GB RAM. The execution time is about 8 minutes for a stereo image of resolution 400×300.

In the following paragraphs, we discuss three popular image editing functions to show the effectiveness of the proposed method. Stereo images are shown as red(left)-cyan(right). More results can be found in the supplementary.

Depth Mapping: This is to modify the depth range of a stereo image. It is typically achieved by adjusting the content disparity of the stereo image. Our depth mapping function is based on linear depth mapping proposed by [11]. The relationship between disparity and depth perceived by the viewer can be described as:

$$Z = \frac{et}{e - s} \tag{8}$$

where e is the interaxial distance set as $65mm$. t is the screen distance from the viewer and is set to $500mm$ in all our experiments. $s = d/\beta$ is the parallax of a matched pixel pair. d is the

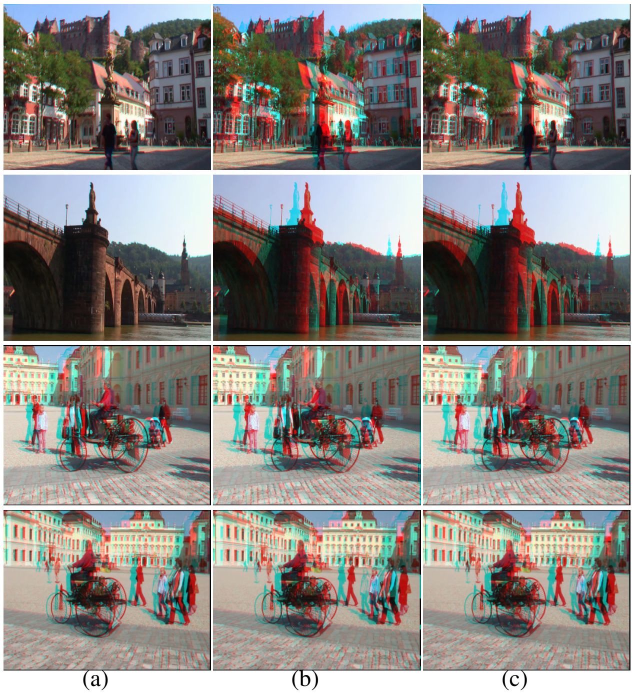

Figure 3: Depth mapping: (a) original stereo image; (b) our result; (c) result from [11]. Target depth range set using $\eta_1 = 0$ and $\eta_2 = 1^o$.

disparity a matched pixel pair and β is the pixel density of the screen. Given the retinal disparity limits η_1 and η_2, the target depth range then becomes (Z'_{min}, Z'_{max}), where $Z'_{min} = et/(e - \eta_1 t)$ and $Z'_{max} = et/(e - \eta_2 t)$.

Fig. 2 shows that our method can enlarge the depth range of the original image in such a way that the small fish becomes nearer to the viewer, without destroying the background depth layer. Fig. 3 compares our method with [11]. The proposed method performs favorably against [11] in two aspects. First, our method can map an original depth range to a much larger depth range, while the editable depth range of [11] is limited by the mesh edge stretching constraint and other constraints. As an example, although both methods are set to the same target depth range in Fig. 3, we can see that the depth range of our result shown in the first and fourth rows of Fig. 3 is much larger than that of [11]; in the fourth row of Fig. 3, the depth distance between the car and the women on the right and the depth distance between the car and the distant men on the left are enlarged more obviously in our result than that of [11]. Second, our method preserves 3D scene structure better. As an example, our method can preserve the shape and the details of the buildings on the right of the images in the first and third rows of Fig. 3 better than [11]. In second row of Fig. 3, our method also preserves the lamppost and the fence on the bridge much better than [11]. In addition, the left and right images of each stereo result show that our method preserves 2D image features much better than [11] too.

Object Depth Adjustment: This is to change the depth of a selected/prominent 3D object. In Fig. 4, we move the man in the middle closer to the viewer. After the operation, our optimization step automatically fills the disoccluded pixels without obvious distortion.

Non-homogeneous Image Resizing: This is to change the aspect ratio of a stereo image non-homogeneously to adapt it to different displays. We compare our method with [1], which is a representative non-homogeneous method for resizing stereo images, as shown in Fig. 5. We can see that our method can preserve the plant and the flower pot much better than [1], and there is no perceptible distortion in our result.

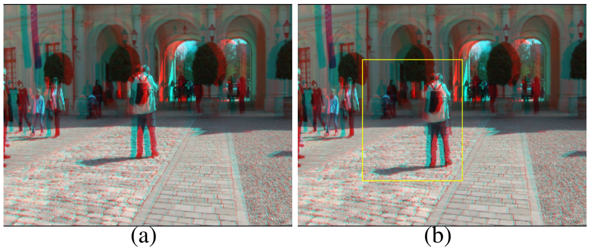

Figure 4: Object depth adjustment: (a) original stereo image; (b) our result.

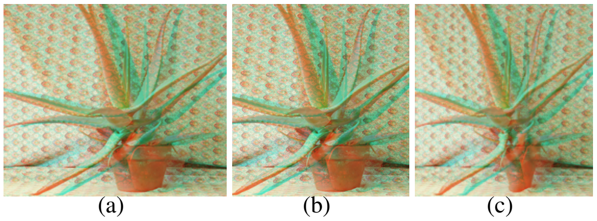

Figure 5: Non-homogenous stereo image resizing: (a) original stereo image; (b) our result; (c) result from geometrically consistent stereo seam carving [1].

5. CONCLUSION AND FUTURE WORK

In this paper, we extend the shift-map method for stereo image editing. The proposed method enforces photo consistence between the left and right images and preserves 3D scene structures well. It can also handle occlusion and disocclusion. We have demonstrated the application of the proposed method on depth mapping, object depth adjustment and non-homogenous image resizing.

The main limitation of the proposed method is the high computational cost, which depends on the size of the search window. We are currently working on a multi-resolution approach to address this problem. On the other hand, we are also exploring more stereo image editing functions based on our framework.

The work presented in this paper was partially supported by a SRG grant from City University of Hong Kong (Project Number: 7002768) and two National Natural Science Foundations of China (Project Numbers: 61033009 and 60970085).

6. REFERENCES

[1] T. Basha, Y. Moses, and S. Avidan. Geometrically consistent stereo seam carving. In *IEEE CVPR*, 2011.

[2] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *IEEE Trans. on PAMI*, 2001.

[3] C.-H. Chang, C.-K. Liang, and Y.-Y. Chuang. Content-aware display adaptation and interactive editing for stereoscopic images. *IEEE Trans. on Multimedia*, 2011.

[4] R. feng Tong, Y. Zhang, and K.-L. Cheng. Stereopasting: Interactive composition in stereoscopic images. *IEEE TVCG*, 2012.

[5] V. Kolmogorov and R. Zabin. What energy functions can be minimized via graph cuts? *IEEE Trans. on PAMI*, 2004.

[6] M. Lang, A. Hornung, O. Wang, S. Poulakos, A. Smolic, and M. Gross. Nonlinear disparity mapping for stereoscopic 3D. *ACM TOG*, 2010.

[7] K.-Y. Lee, C.-D. Chung, and Y.-Y. Chuang. Scene warping: Layer-based stereoscopic image resizing. In *IEEE CVPR*, 2012.

[8] S.-J. Luo, I.-C. Shen, B.-Y. Chen, W.-H. Cheng, and Y.-Y. Chuang. Perspective-aware warping for seamless stereoscopic image cloning. *ACM TOG*, 2012.

[9] Y. Niu, W.-C. Feng, and F. Liu. Enabling warping on stereoscopic images. *ACM TOG*, 2012.

[10] Y. Pritch, E. Kav-Venaki, and S. Peleg. Shift-map image editing. In *IEEE ICCV*, 2009.

[11] T. Yan, R. W. H. Lau, Y. Xu, and L. Huang. Depth mapping for stereoscopic videos. *IJCV*, 2013.