MULTI-VIEW IMAGE SYNTHESIS FROM TWO-VIEW IMAGES USING A CONTENT-PRESERVING WARPING METHOD

† *Yu-Hsiang Huang (黃昱翔)*, † *Yan-Hsiang Huang (黃彥翔)*, † *Tzu-Kuei Huang (黃子魁)*,
‡ *Wei-Chao Chen (陳維超)*, † *Yung-Yu Chuang (莊永裕)*
Dept. of Computer Science and Information Engineering,
National Taiwan University, Taipei
E-mail:†{edwardhw, litleyelow, kuei, cyy}@cmlab.csie.ntu.edu.tw, ‡weichao.chen@gmail.com

ABSTRACT

This paper introduces an automatic warping-based method to synthesize multi-view images from two-view images without requiring depth maps. DIBR is a well-known multi-view synthesis method. However, its success heavily depends on accurate depth maps, which are often still unreliable from only two views with today's technology. Our method finds dense and reliable features as semi-dense stereo correspondences to warp the original binocular views to novel views which satisfy stereoscopic properties while simultaneously preserving the structure of the content. Compared to DIBR, the proposed method can synthesize high-quality multi-view images more efficiently without complex parameter setting. It can be used to convert two-view images taken by binocular cameras into multi-view images so that they can be displayed on autostereoscopic displays.

Keywords: Multi-view Image Synthesis; Autostereoscopy; Content-Preserving Warping;

1. INTRODUCTION

In recent years, 3D multimedia becomes more and more popular. For most today's stereoscopic displays, viewers still need to wear special glasses to watch 3D. It is inconvenient and uncomfortable. In addition, these displays are not suitable for applications which intend to display stereoscopic contents in public space such as 3D advertising billboards. Thanks to autostereoscopic displays, also called glasses-free 3D displays, with them, viewers can enjoy 3D media without wearing annoying glasses. These displays generate three-dimensional effects often by employing lenticular arrays or parallax barriers. As figure 1 shows, the principle is to let eyes see different views at different locations. Thus, the left and right eyes see different views and the viewer perceives 3D.

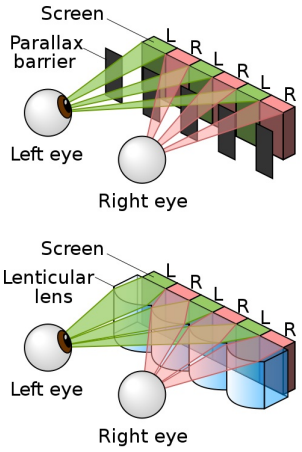


Figure 1: Autostereoscopic displays with a parallax barrier (top) and a lenticular array (bottom). This illustration is from Wikipedia [1]

Although autostereoscopic displays enable glasses-free 3D viewing, they require multi-view contents (normally no less than eight views) rather than two views for stereoscopic displays. Unfortunately, most today's 3D cameras, even professional ones, can only capture two views. Thus, in order to show two-view 3D contents on autostereoscopic displays, it is required to convert them from two-view to multi-view before supplying to the displays. Traditional approaches use stereo

matching methods [3] to find dense depths or disparities from two views, and then apply DIBR methods [4, 5] to synthesize multi-view images [12]. The quality of synthesized views using this kind of methods depends on the accuracy of depth maps. If the depth map is not accurate enough, there will be obvious artifacts at the locations assigned with erroneous depth values. Unfortunately, even with the state-of-art stereo methods, it is still difficult to obtain accurate depth maps automatically and efficiently with only two views.

We propose a new method which uses warping to synthesize novel views without depth maps. First, the method extracts matched feature points between the two input views. Given the parameters of the novel view, the locations of feature points in the novel view are estimated. Next, novel views are synthesized through image warping guided by the estimated feature locations. Additional constraints are added during warping to keep the integrity of the synthesized view by avoiding significant image content distortion. To achieve this efficiently, the input view is divided into a quad mesh. The locations of mesh vertices are optimized with respect to a devised energy function which respects the target feature locations while preserving the content as much as possible. Finally, the novel view is synthesize by warping the input view, guided by the optimal vertex locations.

The rest of this paper is organized as follows. In Section 2, we discuss related work. Section 3 introduces the proposed method. Section 4 presents experiments and comparisons. Finally, we offer our conclusions and describe future work in Section 5.

2. RELATED WORK

Dense Interest Point. To synthesize virtual views by warping, we need to find correspondences. The most popular way for stereo view synthesis [2] is to calculate dense stereo correspondences, such as depth maps or disparity maps. In this paper, we need only sparse stereo correspondences. Thus, feature correspondences between the input images would be sufficient. Standard feature extraction methods, such as SIFT [6] or SURF [10], find good features. However, such methods find very few features in the texture-less regions, and it could cause serious artefacts with our method since those regions could be seriously distorted. Therefore, we chose a dense interest point (DIP) algorithm [16] which combines the advantage of uniform sampling and standard feature extraction: finding features uniformly over the image while maintaining the quality of extracted features.

Content-Preserving Warps. A solution for determining the destination for each pixel in the novel view is to take the feature correspondences as hard constraints. The rest of the pixels are warped according to their neighboring features. However, this method may break the structure of the content seriously, especially along the edges, and thus resulting in visible artifacts. We instead take feature correspondences as soft constraints and to solve the problem by a content-preserving warp [9]. Its advantage is that the warped result better preserves the content.

Grid Line Bending. The smoothness term of Liu et al.'s warping method [9] can preserve the content well. However, it is too strong for our case since the method is devised for video stabilization, in which the input is only slightly deformed. In contrast, our goal is to synthesize virtual views potentially at very different viewing angles and thus more deformation could be necessary. Thus, we use the line bending term for image resizing [17] instead.

3. MULTI-VIEW IMAGE SYNTHESIS

In this section, we describe the proposed multi-view synthesis method in detail. Figure 2 shows the flow chart of our system, which comprises three parts: *semi-dense stereo correspondence*, *virtual camera* and *content-preserving warps*. In the first step, we extract features between the given left-eye and right-eye images. To find the correspondence of each feature, the matching and outlier removal processes are applied. Second, for each virtual camera, we estimate the locations of features in the virtual view by interpolation or extrapolation of original feature coordinates. Finally, we warp the original view according to the estimated feature locations using a content-preserving warping method.

3.1. Semi-Dense Stereo Correspondence

3.1.1. Feature Extraction

In general, there are two types of methods to extract a set of representative points form an image: feature extraction and dense sampling. The latter

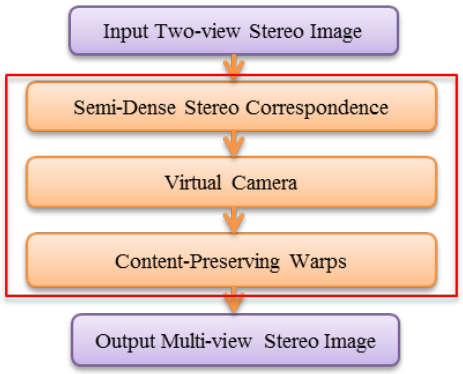


Figure 2: The system flow chart.

is less used because of the tremendous number of sampled points. Nonetheless, for our application, dense sampling has several advantages; it provides better coverage of the image, a stable number of features within the image, and simple spatial relations among features.

In our implementation, we built a scale-space pyramid with 4 octaves. Within each octave, each level of image is blurred using a Laplacian-of-Gaussian (LoG) filter with growing σ's. The number of pyramid levels per octave is 16, which is slightly more than usual. Thus, the initial σ is 1.6 and it is updated with the scale $2^{(1/16)}$. The image is resized by half per octave. Typical feature extraction methods would choose local maximums within any $3 \times 3 \times 3$ neighbourhood as candidates for feature points and then apply non-maximum suppression to retain significant ones. However, in order to achieve dense sampling, the non-maximum suppression is not applied in our method. Thus, for efficiency, we used a larger search area. In the current implementation, the spatial search range is 16-pixel, and the scale search range is 8-octave. That is, our feature extraction method finds the local maximums within all $16 \times 16 \times 8$ neighbourhoods. Note that the property of local maximum is the only criterion for being qualified as features; that is, as long as a point is a local maximum, no matter whether the feature response is high or low, the point is retained as a feature. As the result, the extracted features distribute almost uniformly over the image. Figure 3 shows the dense feature points extracted using this method on the right; the left of the figure shows SURF features for comparisons.

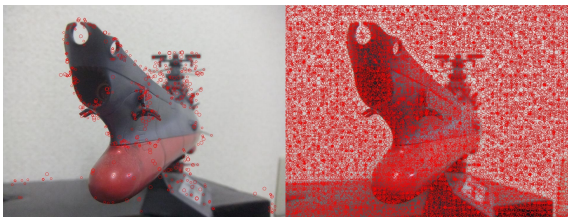


Figure 3: (Left) The features extracted by SURF. (Right) The feature extracted using our dense sampling method. Note that there are much more features than SURF and they are distributed all over the image.

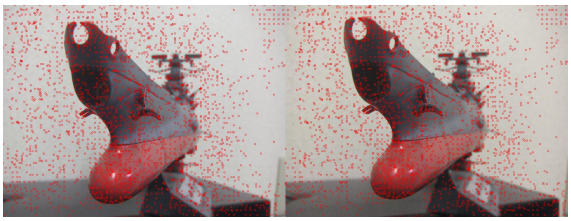


Figure 4: The matched dense interest points using the proposed method on the left and right images.

3.1.2. Feature Matching

For each feature point extracted from the above step, we compute its SURF descriptor. Next, an ANN (Approximate Nearest Neighbour)-based [13] algorithm is employed to find matched features between the two input images. Since the matching result inevitably contains outliers, outlier removal is required for better robustness. Because of the large number of features, for efficiency, we did not use the popular RANSAC method [14]. Instead, we explore the stereo property of the input images and only perform local search for matching features. Figure 4 demonstrates the matched feature points in left-eye and right-eye images.

3.2. Virtual Camera

After extracting and matching semi-dense features, the next step is to estimate where these features should locate in the virtual view. Once we know this information, we can use these feature locations to guide the warp to obtain the full content of the virtual view. We assume that all cameras (including both input real cameras and synthesized virtual cameras) are configured with a parallel setup. That

is, they are arranged along a line called the baseline and all are with the same orientation which is perpendicular to the baseline. We also assume that the distances between two adjacent cameras are all the same.

The right of Figure 5 shows the camera projection model [15]. A 3D point F corresponding to some feature in the scene is projected on the image plane LR of camera R. The x-coordinate of F's corresponding feature on the image plane is M, where $M = BC$. The x-coordinate of the 3D point F in the camera coordinate system is S, where $S = AC$. By the principle of similar triangles, $M/S = BP/AF$. Note that BP/AF is a constant among all these parallel configured cameras. We denote it as K; that is $K = AF/BF$. We can rewrite S as

$$S = M * AF/BF = M * K.$$

The left of Figure 5 describes the relation between features in the virtual views and features in the real views. Assume that the camera R and camera G are both real cameras, while the camera B is the virtual camera at their midpoint. Since The distance between any two adjacent cameras is the same, we have

$$(S_R + S_G)/2 = S_B.$$

We can infer that the following is also true,

$$(M_R + M_G)/2 = M_B.$$

Therefore, we can simply use interpolation and extrapolation to infer the feature points' locations in the virtual views.

3.3. Content-Preserving Warps

Given desired locations of features in the virtual view to be synthesized, for the last part of the proposed method, we first divide the input real view image into $m \times n$ quad grids. Let $\mathbf{V}, \mathbf{E}, \mathbf{G}$ denote grid vertices, grid edges, and grid cells of the original quad mesh, respectively. Also, $\mathbf{V}', \mathbf{E}', \mathbf{G}'$ denote the ones of the mesh after deformation.

3.3.1. Preprocessing

Before optimization, we first compute the saliency of each grid. The higher the saliency of a grid is, the more important a grid is. In our method, we use the squared variance among intensity values of all pixels inside the grid as the grid saliency. Although there are other more complex methods available for computing saliency, we found this simple method is good enough for our application. Next, we compute the best-fit homography by fitting all feature correspondences in a least-squares sense. A global warping based on this homography is applied to warp the input image as an initial guess to the following optimization. We will first describe the data term and smoothness term in the energy function devised for the optimization.

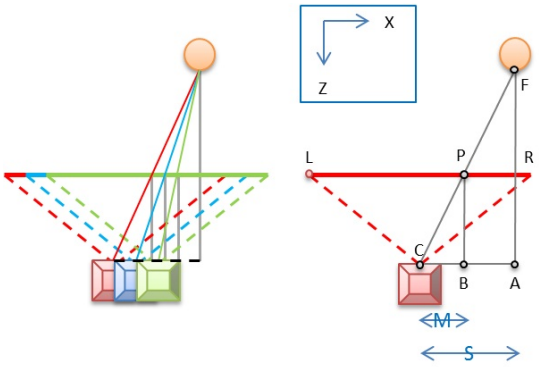


Figure 5: The camera projection model. The red camera is camera R, the blue camera is camera B, and the green camera is camera G.

3.3.2. Data term

To guide the grid cells containing matched features to the desired destination, we need to define the relation between the feature and the grid cell enclosing it. In our method, we represent each feature f by the bilinear interpolation of the four corner vertices v_k of the grid cell g_i enclosing f, where $v_k \in \mathbf{V}$ and $g_i \in \mathbf{G}$. Let w_k denote the weight for the bilinear interpolation for v_k. The data term is then constructed to minimize the $L2$-norm distance between f', the desired location of f computed from the above step, and the point computed by applying w_k to warped grid vertices v_k', the four corners of g's correspondent grid cell g' as follows:

$$E_d = \sum_{\forall f'} \left\| \sum_{k=1}^{4} w_k v_k' - f' \right\|^2.$$

3.3.3. Smoothness term

We use the grid line bending energy as the smoothness term to ensure there is not much content distortion. It is shown as follows:

$$E_s = \sum_{\mathbf{e}\in\mathbf{E},\mathbf{e}'\in\mathbf{E}'} w_{\mathbf{e}} \|s_{\mathbf{e}}\mathbf{e} - \mathbf{e}'\|^2,$$

where $w_{\mathbf{e}}$ is the averaged saliency of the two grid cells that share e as their edge and $s_{\mathbf{e}} = \|e'\|/\|e\|$. This term encourages the edges to remain its orientation throughout the warping. Edges that are shared by high saliency grid are very likely close to prominent contents so they will have larger weights, and vice versa.

However, this term is a non-linear function of e'. This raise the computation complexity of optimization. Wang et al.solved a similar function involving the same smoothness term by using an iterative method [17]. In order to simplify the problem, we replace $\|e'\|$ by $\|\hat{e}\|$, which can be computed from e after preprocessing. This modification works since the initial guess homography are good approximation in most cases. By doing so, this terms can be converted into a linear one.

3.3.4. Optimization and synthesis

Both data term and the smoothness term are combined together to form the total energy function:

$$E = E_d + \alpha E_s,$$

where α is the relative weight between those two energy terms. Since E is a quadratic and sparse linear system, we can solve it by using standard sparse linear solvers. We then synthesize the novel virtual view by applying standard texture mapping algorithm to texture map the original view onto the warped quad mesh.

4. EXPERIMENTAL RESULTS

We collected a set of binocular stereoscopic images from the web for experiments. We used an Alioscopy 3D HD 42" autostereoscopic display to verify the results. This display requires eight views. In addition to two input views, we applied our method to synthesize six virtual views, two extrapolated views on the left of the input left view, two interpolated views between the input views and two views on the right of right views. Figure 6 and figure 7 compare our method with a DIBR method [12] using depth maps generated by a state-of-the-art stereo matching method [3]. Note the depth map is produced by the program without manual intervention, thus there are errors in disparity values, leading visual artifacts in the synthesized images. For example, some artefacts appear in the occlusion region due to the defective inpainting technique, such as the ones shown in Figure 6. Furthermore, the incorrect disparity map would induce some broken region on the right side of the image, such as the ones shown in Figure 7.

We used multi-view data sets of [7], [8], and [11] as ground truth to further evaluate our method. We took view 3 and view 6 as the input left and right views, respectively and applied our method to synthesize view 4 and view 7. We compared the synthesized views with the ground truth in the data set. In general, our method have good results, but could have problem with the regions with prompt depth changes. Additionally, the content in the occlusion region may not be correctly warped since it has no feature correspondence as guidance. However, the artifacts are not very noticeable.

In term of computation time, for traditional approaches, the depth estimation algorithm [3] and DIBR algorithm [12] generally take more than ten minutes combined. The proposed method requires less than a minute on the same hardware. Finally, we show other result of our method in Figure 10.

5. CONCLUSION AND FUTURE WORK

In this paper, we present a method that can synthesize virtual views along the baseline of the input two-view stereo image pair automatically. Different from the majority of today's view synthesis algorithms, our work is a warping-based method that relies on robust semi-dense features rather than brittle depth maps. Thus, the proposed method has the advantages of containing less visual artifacts and being automatic and efficient.

The uniform quad mesh we used in the optimization is independent to the image content. Thus, it is not necessarily consistent to the image content. Thus, even with the help of the content-preserving warp, the results can still have visible distortion when the image content contains strong structures or depth discontinuity. We will employ adaptive

triangle meshes to better matches image content. In addition, we would like to extend the method to stereoscopic videos and explore the possibility of real-time synthesis with the help of GPUs.

ACKNOWLEDGEMENT

This work was partly supported by Taiwan's National Science Council through grants NSC99-2628-E-002-015 and NSC100-2622-E-002-016-CC2.

REFERENCES

[1] Parallax barrier vs lenticular screen, `http://en.wikipedia.org/wiki/File:Parallax_barrier_vs_lenticular_screen.svg`

[2] A. Smolic, P. Kauff, S. Knorr, A. Hornung, M. Kunter, M. Muller, and M. Lang, "Three-Dimensional Video Postproduction and Processing," *Proceedings of IEEE*, Vol. 99, No. 4, pp. 607-625, April 2011.

[3] B.M. Smith, L. Zhang, H. Jin, "Stereo Matching with Nonparametric Smoothness Priors in Feature Space," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 485-492, 2009.

[4] C. Fehn, "A 3D-TV Approach Using Depth-Image-Based Rendering (DIBR,)" *Proceedings of 3rd IASTED Conference on Visualization, Imaging, and Image Processing*, pages 482-487, 2003.

[5] C. Fehn, "Depth-image-based rendering (DIBR), compression, and transmission for a new approach on 3D-TV," *Stereoscopic Displays and Virtual Reality Systems XI. Proceedings of the SPIE*, Vol. 5291, pp. 93-104, 2004.

[6] D.G. Lowe, "Object Recognition from Local Scale-Invariant Features," *Proceedings of the International Conference on Computer Vision*, Vol. 2, pp. 1150-1157, 1999.

[7] D. Scharstein and R. Szeliski, "High-accuracy stereo depth maps using structured light," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Vol. 1, pp. 195-202, 2003.

[8] D. Scharstein and C. Pal, "Learning conditional random fields for stereo," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp.1-8, 2007.

[9] F. Liu, M. Gleicher, H. Jin, and A. Agarwala, "Content-Preserving Warps for 3D Video Stabilization," *ACM Transactions on Graphics (Proceedings of SIGGRAPH)*, Vol. 28, No. 3, p. Article 44, 2009.

[10] H. Bay, A. Ess, T. Tuytelaars, and L.V. Gool, "Speeded-Up Robust Features (SURF)," *Computer Vision and Image Understanding*, Vol. 110, No. 3, pp. 346-359, 2008.

[11] H. Hirschmüller and D. Scharstein, "Evaluation of cost functions for stereo matching," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp 1-8, 2007.

[12] K.-S. Tsai, *Real-Time Preview System for Glass-Free 3D Displays*, Master's thesis, National Taiwan University, 2011.

[13] M. Muja and D.G. Lowe, "Fast Approximate Nearest Neighbors with Automatic Algorithm Configuration," *International Conference on Computer Vision Theory and Applications*, pp. 331-340, 2009.

[14] M.A. Fischler and R.C. Bolles, "Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography," *Communications of the ACM*, Vol. 24, No. 6, pp. 381-395, June 1981.

[15] M. Sonka, V. Hlavac, and R. Boyle, *Image Processing, Analysis & Machine Vision 2nd Edn*, Chapman and Hall, pp. 14, 1995.

[16] T. Tuytelaars, "Dense Interest Points," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2281-2288, 2010.

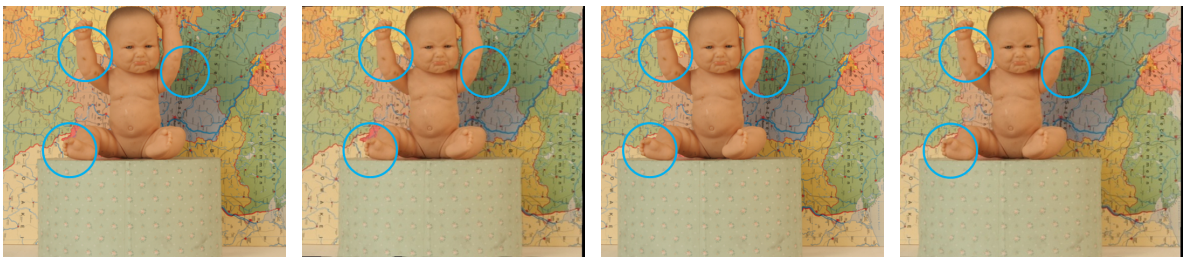
[17] Y.-S. Wang, C.-L. Tai, O. Sorkine, and T.-Y. Lee, "Optimized Scale-and-Stretch for Image Resizing," *ACM Transactions on Graphics (Proceedings of SIGGRAPH)*, Vol. 27, No. 5, p. Article 118, 2008.

(a) Original image (b) Disparity map [3] (c) DIBR method [12] (d) Our method

Figure 6: The highlighted disparity discontinuity in a smooth region cause the duck's head and tail in DIBR's result to distort in a unnatural way. Our result have no such problem.

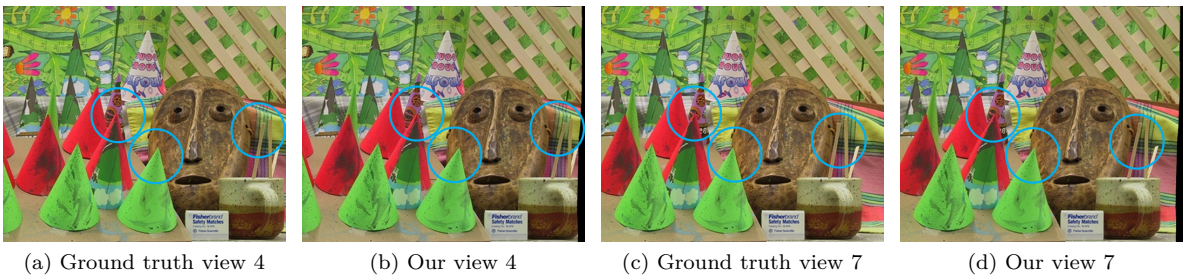
(a) Original image (b) Disparity map [3] (c) DIBR method [12] (d) Our method

Figure 7: The highlighted disparity discontinuity in a smooth region cause a broken poolside in DIBR's result. Our result have no such problem.



(a) Ground truth view 4 (b) Our view 4 (c) Ground truth view 7 (d) Our view 7

Figure 8: A comparison between ground truth and our result. The blue circles point out regions where our results have most visible artifacts.



(a) Ground truth view 4 (b) Our view 4 (c) Ground truth view 7 (d) Our view 7

Figure 9: A comparison between ground truth and our result. The blue circles point out regions where our results have most visible artifacts.

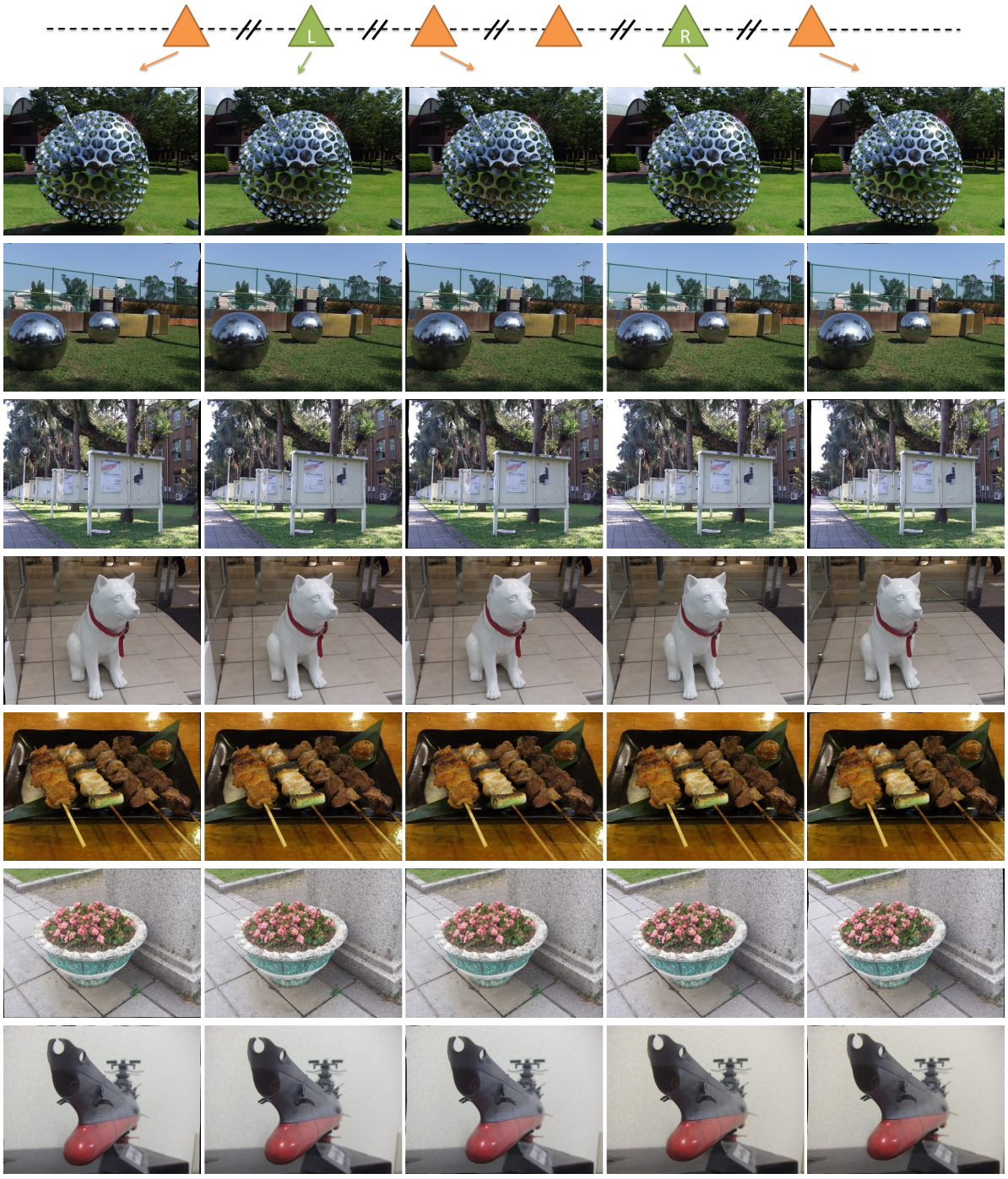


Figure 10: Sample synthesized novel views for seven stereoscopic images. The second and the fourth column show the original left and right views, respectively; the first and the fifth column are the extrapolated virtual views to the left and to the right; and the third column is an interpolated virtual view. Note that the synthesized views do not exhibit any obvious visual artifact. It is the main strength of the proposed warping-based method as it does not reply on brittle depth estimation. In addition, the proposed method has the advantages of being fully automatic and efficient.