Warping-Based Novel View Synthesis from a Binocular Image for Autostereoscopic Displays

Yu-Hsiang Huang* Tzu-Kuei Huang* Yan-Hsiang Huang* Wei-Chao Chen† Yung-Yu Chuang‡
National Taiwan University
Email: {edwardhw, kuei, litleyelow}@cmlab.csie.ntu.edu.tw, weichao.chen@gmail.com†, cyy@csie.ntu.edu.tw‡*

***Abstract*—This paper presents a warping-based method for synthesizing multiple views from a binocular stereoscopic image. Autostereoscopic displays require multiple views while most stereoscopic cameras can only capture two. Popular novel view synthesis methods, such as depth image based rendering (DIBR), often heavily rely on accurate depth maps, which are still difficult to obtain. The proposed method requires neither depth maps nor user intervention. It extracts dense and reliable features. Feature correspondences guide image warping to synthesize novel views while simultaneously maintaining stereoscopic properties and preserving image structures. Compared to DIBR, the proposed method produces higher-quality multi-view images more efficiently without tedious parameter tuning. The method can be used to convert stereoscopic images taken by binocular cameras into multi-view images ready to be displayed on autostereoscopic displays.**

***Keywords*-novel view synthesis; autostereoscopy; warping.**

I. Introduction

Stereoscopic media, such as 3D movies, 3D TV programs and 3D games, have gained more and more popularity in recent years. Accompanied with their fast growth, many stereoscopic displays and binocular cameras have been launched. Most today's stereoscopic displays require viewers to wear special glasses for watching stereoscopic media. It is both inconvenient and less comfortable. In addition, they are not suitable for displaying stereoscopic contents in public, such as for 3D advertising billboards. Thus, autostereoscopic displays, also called glasses-free 3D displays, often made with lenticular arrays or parallax barriers will become the trend for future stereoscopic displays.

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 equipped with special glasses. Unfortunately, most today's stereoscopic cameras, even professional ones, can only capture two views. We do not expect to see cameras equipped with eight lens in the foreseeable future either, because of cost and portability. Thus, it is necessary to convert stereoscopic contents from two views to multiple views before supplying to the displays. Traditional approaches use stereo matching methods [1] to find dense depths or disparities from two views, and then apply DIBR methods [2], [3] to synthesize multi-view images. The quality of synthesized views with this type of methods highly depends on the accuracy of depth maps. If the depth map is not accurate enough, there will be noticeable artifacts at the locations associated with erroneous depth values. Unfortunately, even the state-of-art stereo methods still have difficulties with obtaining accurate depth maps automatically and efficiently from only two views. In addition, most of these methods demand extra information such as additional views and camera parameters, and often take quite an amount of time.

Inspired by recent successes of warping-based methods for various stereoscopic media processing problems, such as disparity remapping [4] and stereoscopic image resizing [5], this paper proposes a method which employs content-preserving warps [6] and line bending energy [7] for novel view synthesis. First, it extracts dense feature matches between the two input views. Given the parameters of the novel view, the locations of features in the novel view are estimated. Next, the novel view is 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 and bending of straight line segments. Because the proposed method only relies on sparse robust features rather than dense error-prone disparity maps, it suffers less from the annoying structure discontinuity artifacts that DIBR methods often exhibit. Experiments show that the proposed method is more effective than DIBR in terms of both time and quality.

II. Related Work

Dense interest points. To synthesize virtual views by warping, we need to find correspondences between views. The most popular way for stereoscopic view synthesis [8] is to calculate dense stereo correspondences, such as depth maps or disparity maps. Standard feature extraction methods, such as SIFT [9] or SURF [10], find good features. However, such methods find few features in the texture-less regions, and this could cause serious artifacts with our method because those regions could be seriously distorted. In this paper, we used semi-dense stereo correspondences. We chose a dense interest point (DIP) algorithm [11] which combines the advantage of uniform sampling and standard feature extraction: finding features uniformly over the image while maintaining the quality of extracted features.

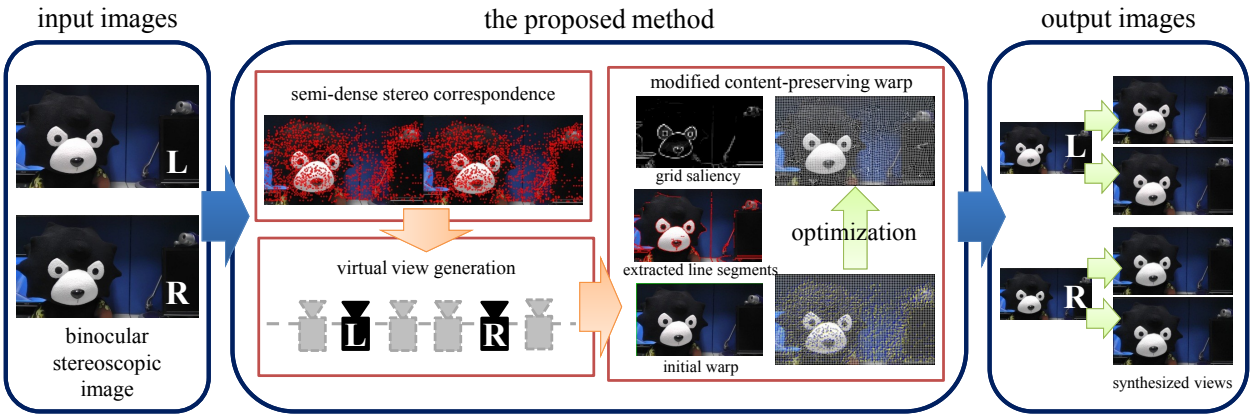

Figure 1. Overview of the proposed multi-view image synthesis method. In the first step, we extract features between two views of the input binocular image pair and find correspondences between features across views through a feature matching process. Next, for each virtual camera, we estimate the locations 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 modified content-preserving warping method which preserves both local shapes and line structures of the content as much as possible.

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 [6]. Its advantage is that the warped result better preserves the content.

Line bending. The smoothness term of Liu et al.'s warping method [6] can preserve the content well. However, it is too strong for our case since their 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 used the line bending term for image resizing [7] instead.

III. Multi-View Image Synthesis

This section describes our method in detail. Figure 1 gives an overview of our system, which comprises three parts: *semi-dense stereo correspondence*, *virtual view generation* and *content-preserving warps*, as detailed in the following.

A. Semi-Dense Stereo Correspondences

Our method requires semi-dense feature correspondences. In general, there are two types of methods for extracting representative interest points form an image: *feature extraction* and *dense sampling*. The latter is less popular because of a tremendous number of interest points. Nevertheless, for our application, dense sampling has several advantages: better coverage of the image, a stable number of features, and simple spatial relations among features.

To combine the advantages of both feature extraction and dense sampling, we follow the idea of a hybrid method called dense interest points (DIP) [11]. First, we construct a Laplacian-of-Gaussian (LoG) pyramid with several octaves, and image is scaled down by the factor of 2 for the next octave. For each octave, an LoG filter of σ varying with its level is applied. We divide the image at each octave into 16×16 non-overlapped patches and find the pixel with the maximal filter response within each patch. The resultant densely sampled points are then checked whether they are local maximums among all of their neighbouring patches at the same level and all of their neighbouring levels. Pixels who are local maximums across levels and octaves are kept and projected back to the finest level. Together, they form a set of dense features. For each feature, we compute its SURF descriptor [10]. Next, we find feature matches of the left and right views. Because of the property of stereoscopic images, matched features should locate roughly on the same scanline. Therefore, we can perform a local search within a small set of neighboring scanlines to both reduce search time and boost accuracy of matches. Figure 2 compares the feature matching results for SURF and DIP. The result of DIP includes both representative matches that also found by SURF and additional matches in the less textured regions such as the window on the left of the image. Such semi-dense feature matches are crucial for our warp-based method. We denote $\mathbf{F}$ as the set of features obtained through the process.

B. Virtual View Generation

After finding semi-dense feature matches, we estimate the locations of these matched features in the virtual view that we want to synthesize. We assume that all cameras, including both real cameras and virtual cameras, are configured with a parallel setting. That is, they are arranged

SURF

DIP

Figure 2. Feature matches with SURF and DIP. DIP has more dense features than SURF, which is crucial for our method.

along a line called the baseline and all are with the same orientation which is perpendicular to the baseline. A virtual view can then be defined by its location on the baseline. For the virtual view to be synthesized, we choose the real view which is closer to the virtual view as the reference image. The locations of the matched feature pairs found in the previous step are then interpolated or extrapolated to calculate their corresponding coordinates in the desired virtual view. We can then use the interpolated/extrapolated features $\mathbf{F}'$ to guide image warping of the reference image to synthesize the virtual-view image.

C. Modified Content-Preserving Warps

From the previous step, we know where the matched features should locate in the virtual view to be synthesized. We divide the reference real-view image into a $m \times n$ quad mesh. Let $\mathbf{V}$ and $\mathbf{E}$ denote vertices and edges of the quad mesh for the reference image, respectively. The goal of content-preserving warps is to find a deformed mesh so that the features move to the desired locations in the virtual view and content structure of the image is preserved as much as possible. Let $\mathbf{V}'$ and $\mathbf{E}'$ be the vertices and edges of the deformed mesh. Note that $\mathbf{E}'$ is defined so that the deformed mesh has a similar edge structure $\mathbf{E}$ as the original mesh. Thus, the only variables we have to solve are vertex positions $\mathbf{V}'$ of the deformed mesh.

1) Preprocessing: The first step is to calculate the saliency value for each quad as its importance. We use the squared variance among intensity values of all pixels within the quad as the quad saliency. More complex methods for saliency calculation exist, but we found this simple method is good enough for our application. Next, we estimate 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 reference image as an initial guess for the following optimization.

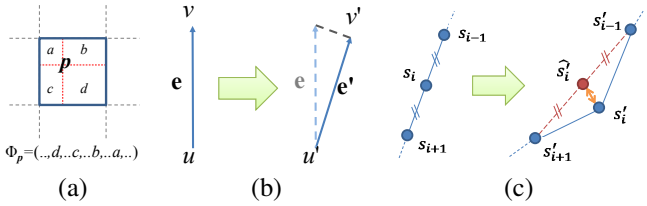

Figure 3. Illustrations of the bilinear weights (a), line bending (b) and line preserving (c).

2) Energy Function: As mentioned, the goal is to find a mesh deformation to guide virtual view synthesis by warping. The deformed mesh should minimize the energy function $E(\mathbf{V}')$ which comprises the following three terms.

Feature correspondence term. This term guides mesh deformation so that the feature locations in the deformed mesh match the desired locations obtained in the Section III-B. Because the feature locations are not the variables to be solved, we have to relate them to the variables, vertex positions $\mathbf{V}'$ of the deformed mesh. Let Φ_p be the bilinear interpolation weight vector for any point p in the reference image; i.e., we have $p = \Phi_p^T \mathbf{V}$. The weight vector Φ_p often only has four non-zero terms which correspond to the weights for the four corners of the quad that p locates within. Figure 3(a) illustrates the setting for Φ_p. After mesh deformation, given the deformed vertices $\mathbf{V}'$, the feature f will move to $\Phi_f^T \mathbf{V}'$. The feature correspondence term ensures that it matches the desired feature location f' by minimizing the L_2-distance between them:

$$E_F = \sum_{\substack{f \in \mathbf{F} \\ f' \in \mathbf{F}'}} \left\| \Phi_f^T \mathbf{V}' - f' \right\|^2. \tag{1}$$

Content coherence term. We use the line bending energy to ensure there is not much content distortion. It requires that the orientation of every edge to be similar before and after mesh deformation, and is defined as:

$$\begin{aligned} E_C &= \sum_{\substack{\mathbf{e}=(u,v) \in \mathbf{E} \\ \mathbf{e}'=(u',v') \in \mathbf{E}'}} w_{\mathbf{e}} \| r_{\mathbf{e}} \mathbf{e} - \mathbf{e}' \|^2 \\ &= \sum_{\substack{\mathbf{e}=(u,v) \in \mathbf{E} \\ \mathbf{e}'=(u',v') \in \mathbf{E}'}} w_{\mathbf{e}} \| r_{\mathbf{e}} (v-u) - (v'-u') \|^2, \end{aligned} \tag{2}$$

where u and v are the vertices at the endpoints of the edge $\mathbf{e}$ before deformation; u' and v' are endpoint vertices of $\mathbf{e}'$ after deformation; $w_{\mathbf{e}}$ is the averaged saliency of the two quads that share $\mathbf{e}$; and $r_{\mathbf{e}} = \|\mathbf{e}'\|/\|\mathbf{e}\|$. With $w_{\mathbf{e}}$, the structures of more important quads will be better preserved. Figure 3(b) illustrates the setting for line bending. Unfortunately, $\|\mathbf{e}'\|$ is non-linear to the variables $\mathbf{V}'$. To simplify the optimization problem, we replace $\|\mathbf{e}'\|$ by $\|\hat{\mathbf{e}}\|$, the length of edge $\mathbf{e}$ after applying homography warping in Section III-C1. This way, $r_{\mathbf{e}}$ becomes a constant during optimization. Since homography provides a very good initial

guess, edges usually do not deform much in lengths. Thus, $\|\hat{\mathbf{e}}\|$ provides a very good approximation to $\|\mathbf{e}'\|$ and the term E_C becomes linear to $\mathbf{V}'$.

Line preserving term. Although the content coherence term preserves the content structure well, we found that object boundaries can still be distorted due to depth discontinuity. The artifacts become especially noticeable if the boundaries are along straight lines. Wang et al. used line preserving term to better preserve straight content edges for their view morphing algorithm [12]. With a similar idea, we add an energy term to refrain straight lines from bending to improve the quality of the synthesized view.

We first extract a set of line segments $\mathbf{L}$ from the reference image using the LSD algorithm [13]. Next, along each line segment, we take samples uniformly between two endpoints. For a line segment l with n samples, we can represent l as $\{s_0^l, s_1^l, ..., s_n^l, s_{n+1}^l\}$, where s_i^l is the i-th sample; and s_0^l and s_{n+1}^l are the endpoints. To preserve straight lines, three consecutive sample points, s_{i-1}^l, s_i^l and s_{i+1}^l, should keep collinearity after mesh transformation. We can measure collinearity by measuring the deviation of $s_i'^l$ from $\hat{s'}_i^l$, the midpoint of $s_{i-1}'^l$ and $s_{i+1}'^l$. Figure 3(c) illustrates the setting. The line preserving term E_L is defined as:

$$E_L = \sum_{l \in \mathbf{L}, \, l=\{s_0^l .. s_{n+1}^l\}} w_l \sum_{i=1}^{n} \|s_i'^l - \hat{s'}_i^l\|^2$$

$$= \sum_{l \in \mathbf{L}, \, l=\{s_0^l .. s_{n+1}^l\}} w_l \sum_{i=1}^{n} \left\| \Phi_{s_i^l}^T \mathbf{V}' - \frac{1}{2}\left(\Phi_{s_{i-1}^l}^T \mathbf{V}' + \Phi_{s_{i+1}^l}^T \mathbf{V}'\right) \right\|^2, \quad (3)$$

where w_l is the weight proportional to l's length before warping. Figure 4 shows the effect of E_L. It is evident that, with E_L, line structures are better preserved in the synthesized view.

3) Optimization and Synthesis: The above three terms are combined together to form the total energy function:

$$E = E_F + \alpha E_C + \beta E_L, \quad (4)$$

where α and β define the relative importance of the last two terms. In our implementation, $\alpha = 5$ and $\beta = 1000$ yield good results in most cases. Since E is in a quadratic form of $\mathbf{V}'$, we can solve it efficiently using standard sparse linear solvers. Finally, we synthesize the novel virtual view by texture mapping the reference image onto the deformed quad mesh. Note that a virtual view is synthesized only from the reference view, the one that is closer to the virtual view. This avoids ghosting artifacts.

IV. Experiments

The input binocular images were either taken by ourselves with a Fujifilm W1 camera or downloaded from public albums. We used an Alioscopy 3D HD 42" autostereoscopic display to verify the results. This display requires eight

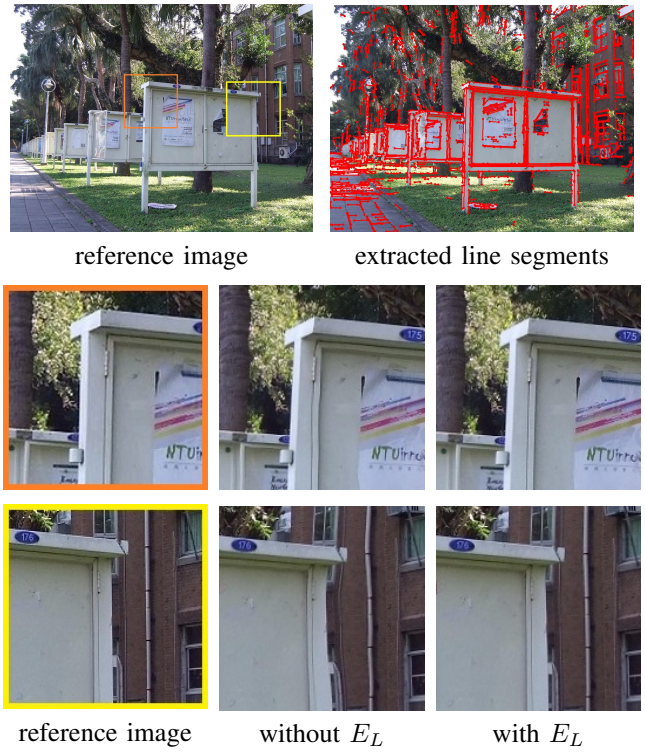

Figure 4. A comparison between the results with and without the line preserving term E_L.

views. In addition to two input views, we synthesized 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 the input right view.

We first compare our method with an inpainting-based DIBR in Figure 5. The depth map was generated by a state-of-the-art stereo matching method [1]. Note that there are errors in disparity values, which lead to visual artifacts in the synthesized images. The incorrect disparity values induce broken and twisted regions, such as the ones shown in Figure 5.

We have also compared our method with the MPEG View Synthesis Reference Software (VSRS) [14]. It requires three views as inputs. On contrary, our method uses only two. We used only two of three views to synthesize virtual views. Figure 6 shows a virtual view generated by our method and VSRS. Compared to VSRS, our method has less ghosting artifacts and other artifacts due to inaccurate depths. In terms of computation time, VSRS took a few minutes for depth map estimation (DERS [15]) and a few seconds for virtual view synthesis. Our method took less than 30 seconds for the whole process. In addition, VSRS requires camera parameters while our method does not need any information other than the input binocular image. Also, VSRS requires three views for high quality results but our method only needs two. In summary, compared with VSRS, our method generates higher quality results more efficiently with less

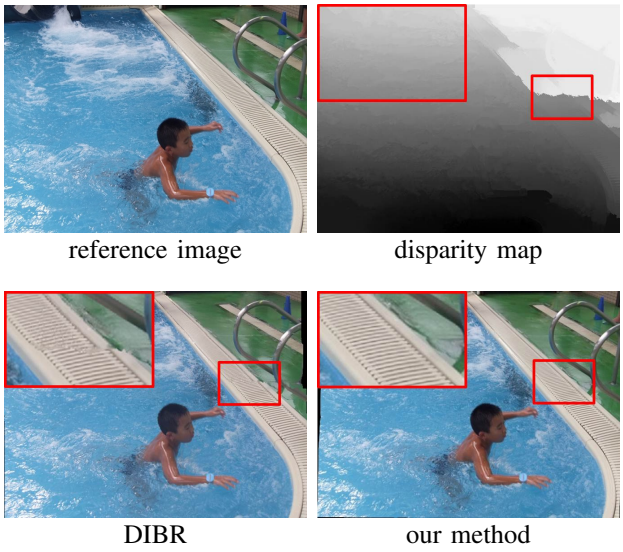

Figure 5. As highlighted, inaccurate disparities cause the broken poolside in DIBR's result. Our result has no such problem.

Figure 6. The first row is a synthesized view using VSRS and our method. The second row provides detail comparisons where the VSRS result contain noticeable artifacts.

images and without tedious parameter settings. Figure 7 shows synthesized views for several stereoscopic images using our method.

V. Conclusion and Future Work

This paper presents a method for synthesizing virtual views along the baseline of the input two-view stereoscopic image pair automatically. Different from popular DIBR methods, rather than brittle depth maps, our method relies on robust semi-dense features. Thus, the proposed method has the advantages of containing less visual artifacts and being automatic and efficient. In the future, we plan to explore the following directions. Using adaptive triangular meshes could better preserve image features. 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 grants NSC100-2628-E-002-009 and NSC100-2622-E-002-016-CC2.

References

[1] B. Smith, L. Zhang, and H. Jin, "Stereo matching with nonparametric smoothness priors in feature space," in *Proceedings of CVPR*, June 2009, pp. 485 –492.

[2] C. Fehn, "A 3D-TV approach using depth-image-based rendering," *Proceedings of 3rd IASTED Conference on Visualization, Imaging, and Image Processing*, pp. 482–487, 2003.

[3] 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.

[4] M. Lang, A. Hornung, O. Wang, S. Poulakos, A. Smolic, and M. Gross, "Nonlinear disparity mapping for stereoscopic 3D," *ACM Transactions on Graphics*, vol. 29, no. 4, pp. 75:1–75:10, 2010.

[5] C.-H. Chang, C.-K. Liang, and Y.-Y. Chuang, "Content-aware display adaptation and interactive editing for stereoscopic images," *IEEE Transactions on Multimedia*, vol. 13, no. 4, pp. 589–601, August 2011.

[6] F. Liu, M. Gleicher, H. Jin, and A. Agarwala, "Content-preserving warps for 3D video stabilization," *ACM Transactions on Graphics*, vol. 28, no. 3, pp. 44:1–44:9, 2009.

[7] Y.-S. Wang, C.-L. Tai, O. Sorkine, and T.-Y. Lee, "Optimized scale-and-stretch for image resizing," *ACM Transactions on Graphics*, vol. 27, no. 5, pp. 118:1–118:8, 2008.

[8] A. Smolic, P. Kauff, S. Knorr, A. Hornung, M. Kunter, M. Mu andller, and M. Lang, "Three-dimensional video postproduction and processing," *Proceedings of the IEEE*, vol. 99, no. 4, pp. 607 –625, April 2011.

[9] D. Lowe, "Object recognition from local scale-invariant features," in *Proceedings of ICCV*, vol. 2, 1999, pp. 1150 –1157.

[10] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (SURF)," *CVIU*, vol. 110, no. 3, pp. 346–359, June 2008.

[11] T. Tuytelaars, "Dense interest points," in *Proceedings of CVPR*, June 2010, pp. 2281 –2288.

[12] H. Wang, M. Sun, and R. Yang, "Space-time light field rendering," *IEEE Transactions on Visualization and Computer Graphics*, vol. 13, no. 4, pp. 697–710, July 2007.

[13] R. von Gioi, J. Jakubowicz, J.-M. Morel, and G. Randall, "LSD: A fast line segment detector with a false detection control," *IEEE TAPMI*, vol. 32, no. 4, pp. 722 –732, April 2010.

[14] ISO/IEC JTC1/SC29/WG11, "View synthesis reference software," May 2009, version 3.0.

[15] ISO/IEC JTC1/SC29/WG11, "Depth estimation reference software," July 2010, version 5.0.

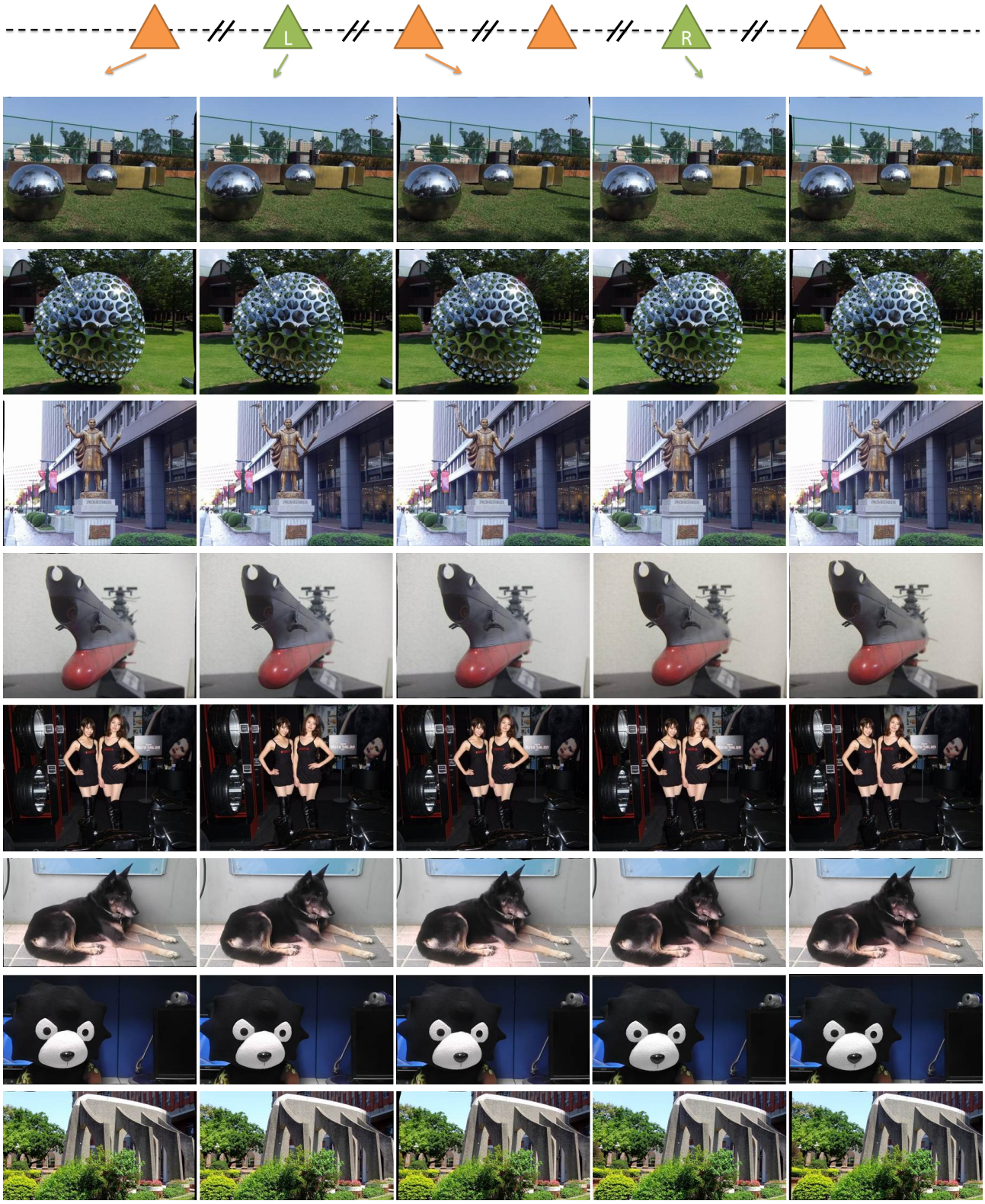

Figure 7. Sample synthesized novel views for several stereoscopic images. Only five of six views are displayed. The second and the fourth column show the original left and right views, respectively; the first and the fifth column are the extrapolated novel views to the left and to the right; and the third column is an interpolated novel view between the left and right views. Note that the synthesized views do not exhibit any obvious visual artifacts. 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.