Spatio-Temporal Coherence for 3-D View Synthesis with Curve-Based Disparity Warping

Hao Wang *, Xiaopeng Zhang *, Hongkai Xiong *

* *Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai 200240, China*

*{sdwfhao123,zxphistory,xionghongkai}@sjtu.edu.cn

***Abstract*—View synthesis is dedicated to generating arbitrary views of the same scene from given inputs. As an alternative to depth-image-based rendering (DIBR), image warping based view synthesis approaches could automatically generate visually plausible virtual views in real-time. Recognizing that existing techniques would lead to temporal incoherence and shape distortions in synthesized videos, this paper proposes a novel video warping algorithm which motion saliency map and global motion from reference views are incorporated into motion-aware constraints to maintain temporal coherence in virtual views. Furthermore, a salient curve based disparity constraint is imposed to prevent shape deformations and avoid possible artifacts. Extensive experiments are validated by visual comparison, which demonstrates that the proposed algorithm outperforms existing warping-based methods.**

***Index Terms*—View synthesis, image warping, spatio-temporal coherence, motion saliency map, salient curve detection**

I. Introduction

Three-dimensional video (3DV) provides the viewers with a realistic 3D impression of the scene, and is considered the next step in the evolution of motion picture formats. With the development of display equipments (e.g. multi-view autostereoscopic displays), glasses-free 3D sensation and motion parallax viewing are enabled in a living room environment. To support the advanced 3DV system, research communities [1] and standardization bodies [2] advocated multi-view video plus depth (MVD) as generic format for 3DV. Such video format consists of a subset M (e.g. 2 or 3) of the N views and corresponding per-sample dense depth for transmission, and additional intermediate views are interpolated using depth-image-based rendering (DIBR) technique [3] at the receiver. Therefore, robustness and quality of virtual views have become decisive factors for viewing experience.

Theoretically, the DIBR algorithm can be utilized to synthesize any virtual perspective views with high accuracy and efficiency. However, the estimated depth maps are usually of insufficient accuracy to enable high quality synthesis, or can't be generated in an automatic and real-time manner. To avoid the restriction imposed by depth estimation, image warping plays as a powerful solution where all positions in one image plane are transformed to positions in a second plane, e.g. video retargeting [4] and video stabilization [5]. Stefanoski et al. [6] developed an image domain warping (IDW) technique as an alternative to depth-based view synthesis, which does not rely on dense depth estimation, but deforms reference view content directly in image space. The reference image with resolution of $W \times H$, is represented as a uniform grid mesh $G = (\mathbf{V}, \mathbf{E}, \mathbf{Q})$ with vertex positions $\mathbf{V}$, edges $\mathbf{E}$ and quads $\mathbf{Q}$. To obtain warped positions $w(v_i)$ for each vertex in the grid, a quadratic energy function $E(w)$ is formulated according to the predefined properties in the synthesized image

$$E(w) \coloneqq \lambda_d E_d(w) + \lambda_s E_s(w) + \lambda_t E_t(w) \tag{1}$$

It consists of three energy terms: a sparse disparity term E_d, a spatial smoothness term E_s, and a temporal coherence term E_t. The disparity term constrains extracted features retaining appropriate disparities after warping, and the spatial smoothness term serves to direct deformations to less salient regions. The temporal coherence term is utilized to preserve smooth and consistent motion in synthesized videos. The warp w can be derived by minimizing $E(w)$ to synthesize the virtual views according to

$$I_{synth}[i, j] \coloneqq \Psi(I, w)[i, j] \coloneqq I(w^{-1}[i, j]) \tag{2}$$

In comparison to DIBR, the warping-based approaches can execute fully automatically and in real-time, without the requirement for dense depth estimation or other error prone processing. Furthermore, the warps are continuous so no holes would occur in the synthesized views, and disoccluded regions are implicitly inpainted by stretching unsalient texture from the neighborhood into the region. This kind of inpainting provides good synthesis results in practice. Nevertheless, the synthesized views would easily suffer flickering artifacts and shape deformations, especially when significant camera or object motions are involved. In addition, the object contours often exhibit double ghost artifacts once sufficient disparity information is not provided for warping process.

To address these problems, this paper proposes a novel view synthesis approach with two contributions, as shown in Fig. 1. As a first contribution, we introduce motion-aware constraints to guide the warp for temporal coherence in synthesized videos. Specifically, the corresponding points in consecutive frames are detected using the global camera motion, and encourage them to undergo the same warp. For the region with moving objects, this constraint could be relaxed by the motion saliency map to avoid possible distortions. As a second

978-1-4799-6139-9/14/$31.00

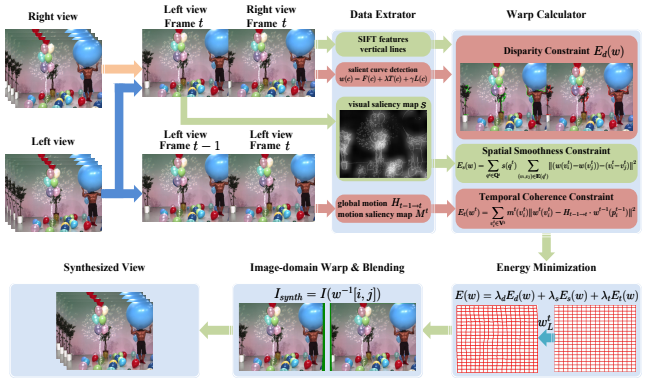


Fig. 1. Overview of the proposed warping-based 3-D view synthesis framework.

contribution, we incorporate a disparity constraint based on salient curve instead of feature points, which can enforce the salient curves to warp into appropriate positions, and reduce the relative artifacts like edge bending.

The remainder of this paper is organized as follows. The motion-aware temporal coherence and salient curve based disparity are presented in Sec. 2 and Sec. 3 respectively. The experimental results are provided in Sec. 4, and Sec. V concludes the work.

II. Motion-aware Temporal Coherence

In essence, the choice of warp is dependent on a trade-off between smooth distortion and good matches [7]. To preserve the temporal coherence in synthesized videos, the energy term in [6] constrains temporally adjacent pixels to warp coherently. Unfortunately, it often fails to guarantee temporal coherence because it assumes that features always remain in the same spatial locations between consecutive frames. Fig. 2 demonstrates an example of shape deformation caused by camera motion. At frame $t-1$, the virtual view is generated from the reference view through an appropriate warping function. Considering the zoomed region in the warping process, the background content (in sky-blue) on left side of the purple balloon has been stretched to fill the disocclusion. This "pull-over" effect would be propagated to the corresponding area in the following frames.

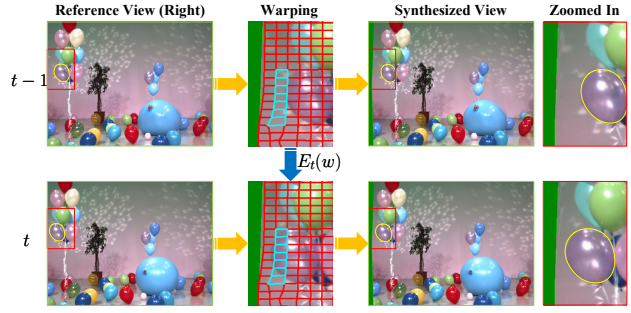


Fig. 2. An example for shape deformation caused by camera motion.

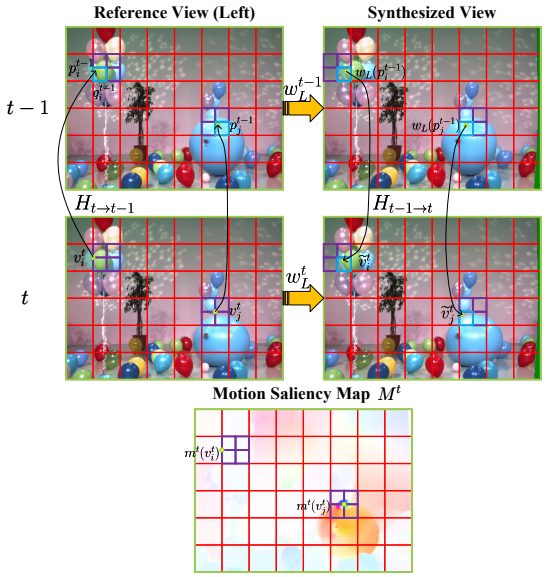


Fig. 3. The motion-aware temporal coherence in the proposed approach.

As illustrated in Fig. 3, a motion-aware constraint is taken into consideration to maintain temporal coherence in synthesized views. For one vertex v_i^t in the mesh grids at frame t, its corresponding position p_i^{t-1} in frame $t-1$ is attained by global motion compensation. Here, we demonstrate the camera motion by a homography model, and estimate it through a SIFT-feature based descriptor [8]. In turn, we can obtain position p_i^{t-1} based on the relationship: $p_i^{t-1} = H_{t\to t-1}\cdot v_i^t$, where $H_{k\to l}$ represents the inter-frame transformation from frame k to frame l. It is reasonable to assume that the virtual view shares the same camera motion as the reference one, and the warping positions of v_i^t and p_i^{t-1} should be constrained to maintain the relationship, i.e., by minimizing $\|w^t(v_i^t) - H_{t-1\to t}\cdot w^{t-1}(p_i^{t-1})\|^2$.

Since p_i^{t-1} might not be a grid vertex, the constraint cannot be applied to p_i^{t-1} straightforward. To solve it, we represent p_i^{t-1} as a bilinear interpolation of the four vertices that enclose it as follows:

$$p_i^{t-1} = \sum_{v_k^{t-1}\in \mathbf{V}(q_i^{t-1})} \alpha_k \cdot v_k^{t-1} \tag{3}$$

where q_i^{t-1} is the corresponding quad that contains the spatial location of p_i^{t-1}, $\mathbf{V}(q)$ represents the vertex set of quad q, and α_k denotes the relative bilinear interpolation coefficients. Thus, the warping position $w^{t-1}(p_i^{t-1})$ can be achieved as follows:

$$w^{t-1}(p_i^{t-1}) = \sum_{v_k^{t-1}\in \mathbf{V}(q_i^{t-1})} \alpha_k \cdot w^{t-1}(v_k^{t-1}) \tag{4}$$

Note that the global motion constraint is based on the assumption that the corresponding positions across frames, v_i^t and p_i^{t-1}, represent the same 3-D point in the scene. It works well for static objects or backgrounds. However, the mismatch usually occurs between the corresponding points from dynamic foreground, e.g. v_j^t and p_j^{t-1} in Fig. 3, as foreground objects have their own motion independent of camera motion. For the region with moving objects, we propose to

relax the temporal constraint and assign more freedom to the relative warping by introducing the motion saliency map. As [9], we obtain the motion saliency map as the difference between the local optical flow [10] and the global background motion. Fig. 3 shows an illustration of the motion saliency map M^t with respect to frame t, where hue indicates orientation and saturation indicates magnitude. Here, the weight $m^t(v_i^t)$ is utilized to adaptively tune the temporal constraint according to the corresponding saliency value:

$$m^t(v_i^t) = \min\{\frac{N}{\sum_{p_k^t \in \mathcal{N}(v_i^t)} M^t(p_k^t)}, 10\} \tag{5}$$

where $\mathcal{N}(v)$ represents the square window of 7×7 pixels centering at vertex v, $M^t(p)$ corresponds to the magnitude value of motion saliency at position p, and N is the number of pixels in $\mathcal{N}(v)$. Thus, the vertices with local motion (e.g. v_j^t) would be assigned to smaller temporal coherence weight than static vertices (e.g. v_i^t), and get relaxed from temporal constraint to avoid possible distortions.

We utilize both global and local motion to obtain the following temporal constraint:

$$E_t(w^t) = \sum_{v_i^t \in \mathbf{V}^t} m^t(v_i^t) \| w^t(v_i^t) - H_{t-1 \to t} \cdot w^{t-1}(p_i^{t-1}) \|^2 \tag{6}$$

III. Salient Curve Based Disparity

For image warping based approaches, the sparse disparities, which are estimated from the Data Extractor in Fig. 1, are critical for 3-D perception of viewers. The traditional SIFT features and vertical edges for disparity estimation can not represent salient curves with flexible shape, and artifacts would obviously occur around object contours, as shown in Fig. 4. To solve the problem, an disparity based on salient curves is imposed on the disparity constraint.

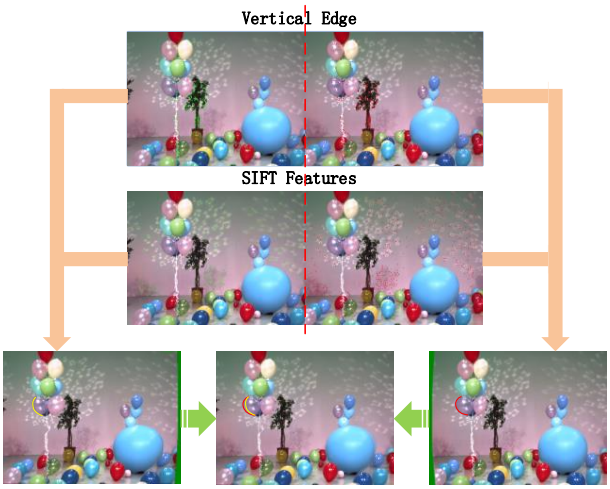
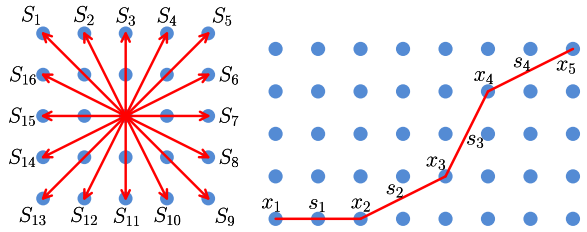


Fig. 4. An example for double ghosting artifacts due to lack of disparity information around the purple balloon.

A. Curve representation

We adopt the model in [11] for salient curves representation. The basic elements of curves are depicted as 16 oriented segments $\mathcal{S} = \{S_1, S_2, \ldots, S_{16}\}$, as shown in Fig. 5 (a). Each segment connects two adjacent points in a curve, and only depends on their relative positions. Thus, a curve can be represented as a sequence of points $c = (x_1, x_2, \ldots, x_n)$, where x_i denotes the coordinate of the i-th point in the curve. Equivalently, in a segment-based form: $c = (s_1, s_2, \ldots, s_{n-1})$, where s_i is one of the oriented segments calculated by $s_i = x_{i+1} - x_i$. Fig. 5 (b) illustrates the two representations of one curve for $n = 5$.



(a) oriented segments (b) curve representation

Fig. 5. An example of curve representation.

B. Curve detection

As human visual system is sensitive to high frequency components, salient curves are connected paths that exhibit intense grayscale variations in an image. In addition, the curves with smooth shape and sufficient length would attract more attention. Hence, the quality of a curve could be evaluated as:

$$w(c) = F(c) + \lambda T(c) + \gamma L(c) \tag{7}$$

The intensity term

$$F(c) = \frac{1}{n} \sum_{i=1}^{n} \sqrt{I_x(x_i) + I_y(x_i)} \tag{8}$$

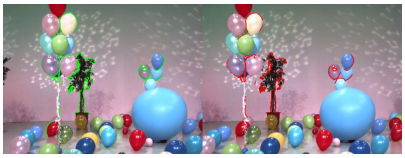
measures the boundary response of the points in curve c, where $I_x(x_i)$ and $I_y(x_i)$ are the gradient of the image with respect to x and y. The smoothness term $T(c)$ indicates the orientation consistency of curve c by measuring the difference between two adjacent oriented segments as:

$$T(c) = \frac{1}{n-2} \sum_{i=2}^{n-1} \exp(-|s_i - s_{i-1}|) \tag{9}$$

The length term $L(c) = \log(n)$ ensures salient curves with sufficient length.

Salient curves with large weights and sufficient length are extracted by maximizing Eq. (7). In turn, the disparities of detected curves are estimated by the Lucas-Kanade algorithm, and incorporated into the disparity constraints E_d. Fig. 6 (a) shows the disparities based on salient curves, where one can observe that sufficient disparities around the purple balloon are provided for warping process. Fig. 6 (b) shows the synthesized frame, where salient shape distortions (e.g. double ghosting

or edge bending artifacts in Fig. 4) are eliminated due to the availability of the disparities.



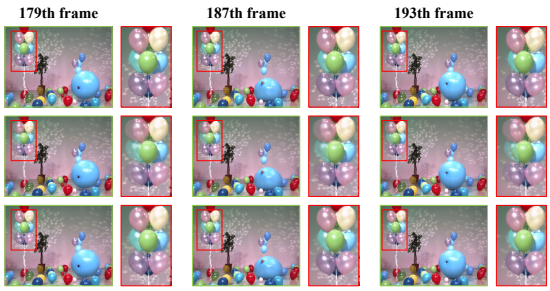
(a) disparities based on salient curve

(b) synthesized view

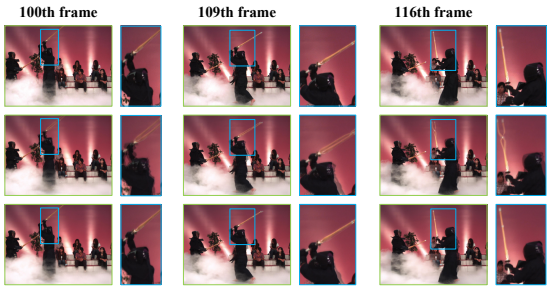
Fig. 6. The disparity based on salient curves and its synthesized effect.

IV. Experimental Results

The proposed algorithm has been compared with the state-of-the-art view synthesis approach [6]. To evaluate the performance of spatial and temporal coherence, the *Balloons* and *Kendo* test sequences (both with the resolution of 1024×768) from Nagoya University are chosen for their fast motion and flexible shapes. Without loss of generality, the middle view 3 is synthesized from view 1 and view 5. We solve the warp with a mesh resolution of 180×100 following the same setting in [6] for stereoscopic high-definition sequences. The synthesized effect is provided in Fig. 7, along with original frames.



(a) The *Balloons* sequence and ghosting artifacts



(b) The *Kendo* sequence and edge bending artifacts

Fig. 7. Comparison of different view synthesis schemes. The first row is the original frames, the second row from [6], the last row from the proposed approach.

From Fig. 7, it can be seen that the proposed scheme exhibits a high quality of rendered views and eliminates the artifacts in [6]. Specifically, Fig. 7 (a) shows the ghosting artifact (second row) caused by global camera motion, and the balloons always appear flickering effect when playing the synthesized video. From the last row in Fig. 7 (a) we can see, our proposed algorithm removes these artifacts using the motion-aware temporal coherence, and the video is more fluent and comfortable for viewing. In Fig. 7 (b), no disparity information is extracted with respect to the "Shinai" and thus severe edge bending artifact is present in synthesized view (second row). With the salient curve based disparity constraints in our proposed scheme, sufficient disparities are provided and the "Shinai" is preserved much better as shown in the last row.

V. Conclusion

In this paper, we propose a novel view synthesis approach with the motion-aware temporal coherence and salient curves based disparities. In contrast to conventional approaches, it is capable of reducing shape deformations and preserving temporal consistency in the videos with fast motion. It can enforce the salient curves to warp into appropriate positions, and reduce the relative artifacts like edge bending. Experimental results show that the synthesized views of the proposed algorithm provide high quality and less artifacts.

Acknowledgment

The work was supported in part by the NSFC, under grants U1201255, 61271218, and 61228101.

References

[1] K. Muller, P. Merkle, and T. Wiegand, "3-D video representation using depth maps," *Proceedings of the IEEE*, vol. 99, no. 4, pp. 643-656, Apr. 2011.

[2] A. Smolic, K. Muller, P. Merkle, and A. Vetro, "Development of a new MPEG standard for advanced 3D video applications," in *Proc. International Symposium on Image and Signal Processing and Analysis*, pp. 400-407, Sep. 2009.

[3] C. Fehn, "Depth-image-based rendering (DIBR), compression, and transmission for a new approach on 3D-TV," *Proc. SPIE*, vol. 5291, pp. 93-104, May 2004.

[4] P. Krähenbühl, M. Lang, A. Hornung, and M. Gross, "A system for retargeting of streaming video," *ACM Trans. Graph.*, vol. 28, no. 5, 2009.

[5] F. Liu, M. Gleicher, H. Jin, and A. Agrawala, "Content-preserving warps for 3D video stabilization," *ACM Trans. Graph.*, vol. 28, no. 3, 2009.

[6] N. Stefanoski, O. Wang, M. Lang, P. Greisen, S. Heinzle, and A. Smolic, "Automatic view synthesis by image-domain-warping", *IEEE Trans. Image Processing*, vol. 22, no. 9, pp. 3329-3341, Sep. 2013.

[7] C.A. Glasbey, K.V. Mardia, "A review of image warping methods", *Journal of Applied Statistics*, vol. 25, pp. 155-171, 1998.

[8] B. Chen, K. Lee, W. Huang, and J. Lin, "Capturing intention-based full-frame video stabilization," *Computer Graphics Forum*, vol. 27, no. 7, pp. 1805-1814, Jul. 2008.

[9] Y. Niu, F. Liu, X. Li, and M. Gleicher, "Warp propagation for video resizing," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, San Francisco, CA, USA, pp. 537-544, Jun. 2010.

[10] C. Liu, "Beyond Pixels: Exploring New Representations and Applications for Motion Analysis," *Ph. D Thesis*, Massachusetts Institute of Technology, May 2009.

[11] P. Felzenszwalb and D. McAllester, "A min-cover approach for finding salient curves," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition Workshop*, Jun. 2006.