360° VIDEO STITCHING FOR DUAL FISHEYE CAMERAS

I-Chan Lo, Kuang-Tsu Shih, and Homer H. Chen

National Taiwan University
E-mail: {f05942036, dafeishih, homer}@ntu.edu.tw

ABSTRACT

Back-to-back dual fisheye camera configuration offers a cost-effective 360° video solution. However, the inherent parallax of this camera configuration and the unmatched scene along the seam between the two camera views present a great challenge to video stitching. To address the jittering and fragmented visual appearance issues, we present in this paper a robust method that preserves the geometric structure of the scene and enhances the stableness of the video along the temporal dimension. The proposed video stitching method entails a mesh deforming operation prior for the minimization of the geometric distortion and an adaptive seam carving operation that generates optimal spatial and temporal alignment. Experimental results show that our method can produce 360° videos without jitter and ghost artifact.

Index Terms**—**Fisheye lens camera, video stitching, parallax, image warping, seam carving, 360° video.

1. INTRODUCTION

Most videos have been captured with a narrow field of view. In contrast, panoramic photography breaks the constraint and greatly extends the field of view by combining multiple images captured by a single camera from different view angles or by a camera array. A properly stitched 360° video brings viewers a completely immersive visual experience for virtual reality applications. It is expected that the demand for 360° video will increase rapidly in fields such as entertainment, education, medical surgery, and gaming. Robust stitching holds the key to stable and seamless 360° videos for these applications.

In this paper, we propose a method to stitch videos captured by dual fisheye cameras as illustrated in Fig. 1. Such camera configuration is found in many commercial products [1]–[3]. However, existing stitching algorithms often generate 360° video with ghost or flickering artifact in the overlapping region of the two camera views because of the following two reasons. First, the parallax of the two fisheye cameras creates a disparity between corresponding features in the overlapping region. Second, the fisheye image distortion introduces unmatched scene along the seam in the overlapping region.

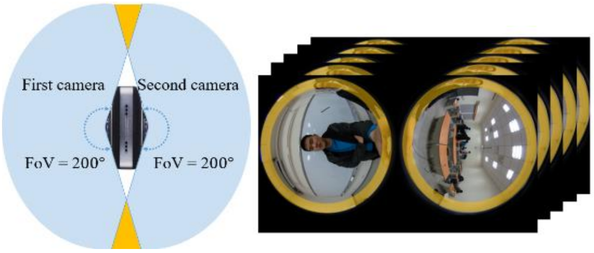


Fig. 1. Configuration of a pair of 200°, back-to-back fisheye cameras. (a) Illustration of the overlapping field of view between the two cameras. (b) A typical dual fisheye image stack captured by the pair of fisheye cameras. The overlapping region is highlighted by a yellow ring on each image.

Based on the observation that equirectangular transformation preserves the magnitude of object motion in the overlapping region, we propose a robust prior for video alignment, which is a critical step for video stitching. The proposed prior takes the patch-wise object motion in the overlapping region into consideration and guides the objective function to preserve the spatiotemporal consistency of the video content. Also, a novel adaptive seam carving method is proposed to optimize the alignment quality by minimizing the spatiotemporal seam cost across consecutive frames. It adaptively updates the seam path according to the video content to prevent flickering artifact.

In short, our stitching method takes the fisheye videos captured by a pair of back-to-back dual fisheye cameras as input and produces a stable and seamless 360° video. The strength of our method is two-folded. First, it preserves spatial continuity of the scene and hence does not produce fragmented appearance or ghost artifact. Second, it retains the temporal continuity of the scene and is free from jittering or flickering.

2. RELATED WORK

We classify existing panoramic video stitching method for 360° video into two categories: The first category takes rectilinear video data as input [4]–[6], and the second category takes fisheye video data as input [7].

Perazzi et al. produced a panoramic video by taking rectilinear video data as input. The video data were captured by an unstructured camera array [4] consisting of a number of cameras assembled in random orders. The resulting video has an enlarged field of view. Lin et al. proposed a video stitching method for hand-held cameras. Their method

978-1-5386-6249-6/19/$31.00 ©2019 IEEE

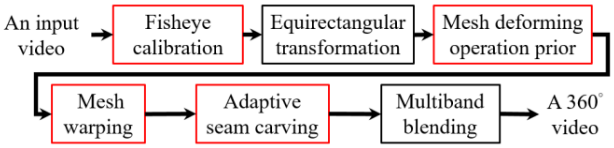


Fig. 2. A block diagram of the proposed method.

synthesizes the output video by preserving line structure [5]. Jiang and Gu also proposed a solution for video stitching using multiple cameras [6]. However, because of the use of traditional feature descriptors (such as SIFT [8] and SURF [9]) for image warping, their method is fraught with difficulties in dealing with fisheye distortion.

To address the issue associated with fisheye distortion, Ho et al. employed rigid moving least squares to stitch the video [7]. However, this method requires the controlling points to be redefined before alignment. Furthermore, it only works for small parallax.

3. PROPOSED METHOD

Fig. 2 shows the block diagram of our proposed method. There are four main steps (marked in red) designed for 360° video stitching: fisheye calibration, mesh deforming operation prior, mesh warping, and adaptive seam carving. The details of each step are described in this section.

3.1 Fisheye calibration

In an ideal $r-\theta$ fisheye projection model, the distance r from the center of the fisheye image to a pixel is linearly proportional to the latitude of the corresponding incident angle θ of the ray originated from the scene to the center of the fisheye lens,

$$r = f\theta \tag{1}$$

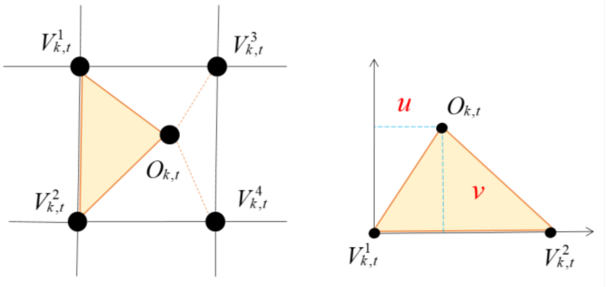
where f denotes the focal length of the fisheye lens. However, the fisheye lens may not be ideal in practice. Therefore, we extend the model by using a quartic polynomial with coefficients a_1, a_2, and a_3 as follows:

$$r + a_1r^2 + a_2r^3 + a_3r^4 = f\theta. \tag{2}$$

The fisheye image calibration is performed by back-projecting the fisheye image onto the surface of a unit sphere. After the calibration, an equirectangular transformation is applied to rectify the fisheye image.

3.2 Mesh deforming operation prior

In this step, the operation is performed on the overlapping (or common) image region of a pair of transformed fisheye images. Let I_t be the target frame captured by the first camera and I'_t be the corresponding reference frame captured by the second camera at time t. Note that a video frame is referred to by the time it is captured hereafter. We divide the target image I_t into an $m\times n$ uniform mesh and denote the four vertices of the k-th patch of the mesh by $V^j_{k,t}$, j=1,⋯,4. We then calculate the magnitude of the motion of each patch as the mesh warping operation prior of the patch. Since the equirectangular transformation preserves the rectilinear property of the overlapping region, the prior $p_{k,t}$ of the k-th patch between frames $t-1$ and t is obtained by



(a) (b)

Fig. 3. An example mesh patch and the linear dependency of vertices. (a) The patch is further divided into four triangles with the centroid as the common vertex. (b) The vertex $O_{k,t}$ can be represented as a linear combination of the vertices $V^1_{k,t}$ and $V^2_{k,t}$.

$$p_{k,t} = \frac{1}{D_k}\left\|\sum_{l=1}^{D_k}(P^l_{k,t} - P^l_{k,t-1})\right\|, \tag{3}$$

where $P^l_{k,t-1}$ and $P^l_{k,t}$ denote the coordinates of dense corresponding points [10] of the patch between I_{t-1} and I_t, respectively, and D_k is the number of corresponding points of the k-th patch. If there is no correspondence, $p_{k,t}$ is zero. The purpose of the mesh warping operation prior is to guide the objective function of mesh warping. If there is object motion in the patch, encourage the vertices of the patch to deform; otherwise, force the vertices to retain the current state.

3.3 Mesh warping

We apply a feature extraction method [11] to find the corresponding feature points $F_{k,t}$ and $F'_{k,t}$ of the k-th mesh patch between I_t and I'_t. In order to have an accurate image alignment, an objective function that combines a feature term, a structure term, and a temporal coherence term is adopted to warp the target image I_t and align it with the reference image I'_t. The deformed mesh vertices $\{\hat{V}_{k,t}\}_{k=1}^{mn}$ of I_t is obtained by minimizing the objective function.

Let $E^f_{k,t}$ denote the energy of deforming the k-th mesh patch for feature alignment. The purpose of the feature term is to minimize the error of the aligned feature points $\hat{F}_{k,t}$ and the corresponding feature points $F'_{k,t}$. Define

$$E^f_{k,t}(\hat{V}_{k,t}) = p_{k,t}\left\|\hat{F}_{k,t}(\hat{V}_{k,t}) - F'_{k,t}\right\|^2, \tag{4}$$

where $\hat{F}_{k,t}$ is obtained by a bilinear interpolation of the four deformed vertices $\hat{V}_{k,t}$ with the sum of coefficients equal to 1. That is, $\hat{F}_{k,t} = w^T_{k,t}\hat{V}_{k,t}$.

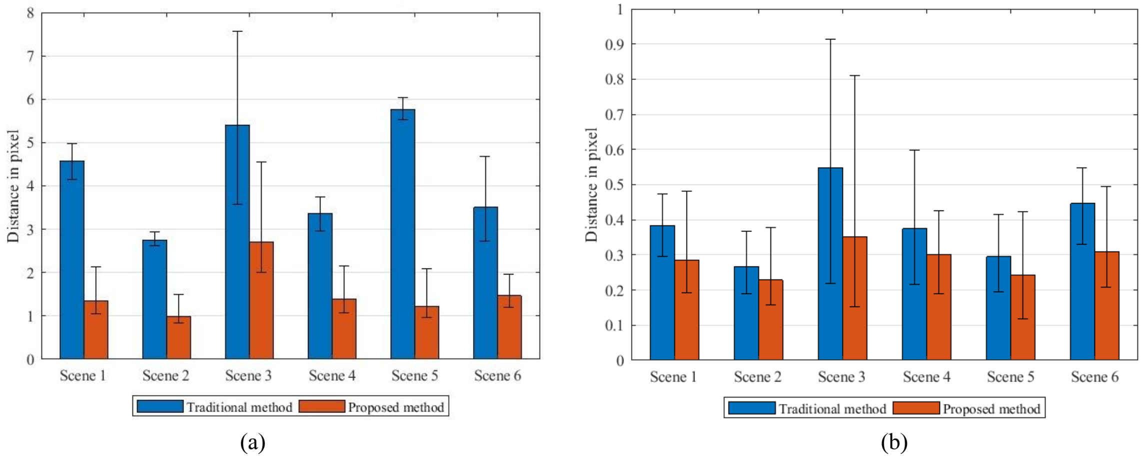
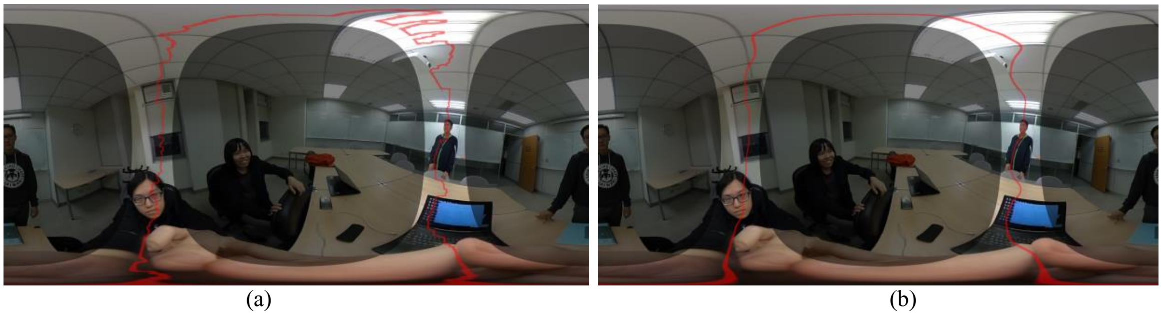


Fig. 4. Performance comparison. (a) Average feature alignment error. (b) Average vertex displacement between consecutive frames.



(a) (b)

Fig. 5. The seam (marked in red) of an example scene. (a) Previous method [11]. (b) Our proposed method. Note that the bright area outside the rounded squares is the overlapping region.

The structure term denoted by $E^s_{k,t}$ describes the energy of preserving the geometric structure of the *k*-th mesh patch. To compute $E^s_{k,t}$, we divide each mesh patch into four triangles. Each triangle consists of two neighboring vertices and the centroid $O_{k,t}$ of the patch, which is defined by the sum of the coordinates of all pixels in the patch weighted by the luminance of the pixels. Afterward, a similarity transformation is applied to every triangle to minimize the mesh deforming distortion. We can represent the centroid as a bilinear combination of the four vertices of the patch. That is, $O_{k,t} = C^T_{k,t} V_{k,t}$ (see Fig. 3(a)), where $C_{k,t}$ denotes the bilinear combination coefficients in the matrix form. As an illustration, the linear dependency between $V^1_{k,t}$, $V^2_{k,t}$ and $O_{k,t}$ is shown in Fig. 3(b), where *u* and *v* are the two parameters that characterize the structure of the triangle defined by the three vertices [12],

$$u = (O_{k,t} - V^1_{k,t}) \cdot (V^2_{k,t} - V^1_{k,t}) \Big/ \left\| V^2_{k,t} - V^1_{k,t} \right\|, \tag{5}$$

$$v = \left\| (O_{k,t} - V^1_{k,t}) \times (V^2_{k,t} - V^1_{k,t}) \right\| \Big/ \left\| V^2_{k,t} - V^1_{k,t} \right\|. \tag{6}$$

Likewise, the warped centroid $\hat{O}_{k,t}$ can be represented as a linear combination of the warped vertices $\hat{V}^1_{k,t}$ and $\hat{V}^2_{k,t}$:

$$\hat{O}_{k,t} = \hat{V}^1_{k,t} + u(\hat{V}^2_{k,t} - \hat{V}^1_{k,t}) + vR_{90}(\hat{V}^2_{k,t} - \hat{V}^1_{k,t}), \tag{7}$$

where

$$R_{90} = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}. \tag{8}$$

The structure term $E^s_{k,t}$ aims to preserve the similarity for the four triangles in each patch and is in the form

$$E^s_{k,t}(\hat{V}_{k,t}) = p_{k,t} \left\| \hat{O}_{k,t}(\hat{V}_{k,t}) - O_{k,t} \right\|^2. \tag{9}$$

The temporal coherence term $E^c_{k,t}$ is introduced to preserve the smoothness in the temporal dimension by encouraging a similarity transformation between the meshes of two consecutive video frames. It takes the form

$$E^c_{k,t}(\hat{V}_{k,t}) = \tilde{p}_{k,t} \left\| \hat{O}_{k,t}(\hat{V}_{k,t}) - \hat{O}_{k,t-1} \right\|^2, \tag{10}$$

where $\hat{O}_{k,t-1}$ and $\hat{O}_{k,t}$ are the warped centroids of the patch in frames $t-1$ and *t*, respectively, and $\tilde{p}_{k,t} = \max_k p_{k,t} - p_{k,t}$ is the complement of $p_{k,t}$. If there is no object motion in the patch, $\tilde{p}_{k,t}$ encourage vertices to retain the state.

The objective function optimizes the overall energy functions using the associated weights *α* and *β*, as follows:

$$E_t = \sum_{k=1}^{mn} \left(E^f_{k,t} + \alpha E^s_{k,t} + \beta E^c_{k,t} \right). \tag{11}$$

We minimize E_t using a sparse linear solver [13] and apply texture mapping to render the warped target image $\hat{I}_t$.

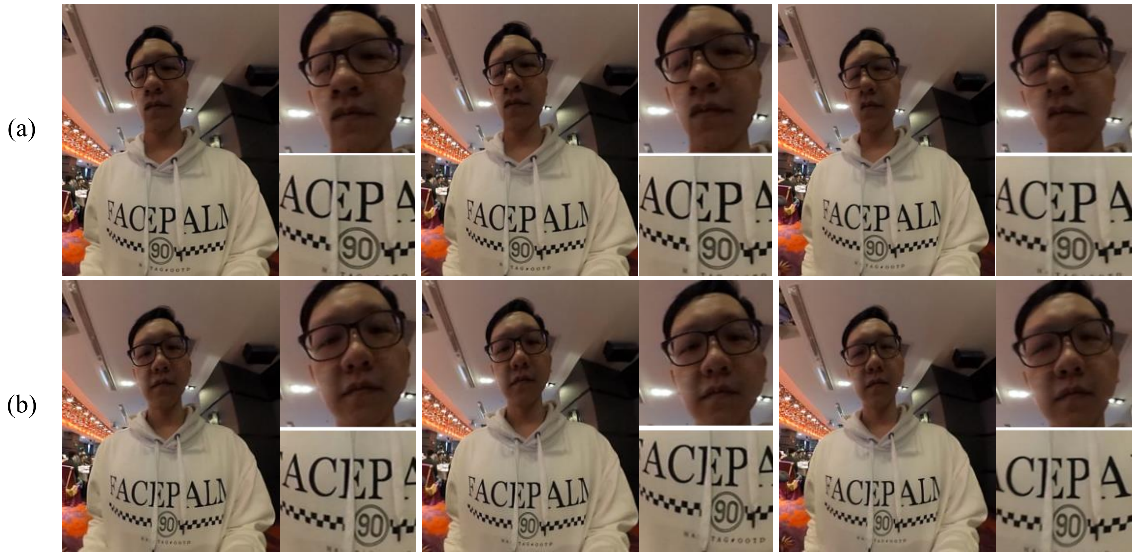


Fig. 6. Close-up views of three stitched frames for comparison. (a) Traditional method generates fragmented video frames. (b) Our method generates seamless and stable video frames.

3.4 Adaptive seam carving

A pair of aligned video frames in an image stack is merged using seam carving and multi-band blending. A seam denoted by S_t is an 8-connected path of pixels from right to left in frame t (see Fig. 5). Let $s_{i,t}$ be a pixel of the path. Then, $S_t = \{s_{i,t}\}_{i=1}^{L}$, where L is the length of the seam. We extend the seam carving method [11], [12] to video by performing a least square minimization on the derivative of S_t to achieve spatial-temporal smoothness. Define image energy function by $e(I) = |\partial I/\partial x| + |\partial I/\partial y|$ to guide seam carving. Then the optimal seam S_t^* for frame t is obtained by minimizing the spatial-temporal cost

$$\sum_{i=1}^{L}\left\| e(\hat{I}_t\,(s_{i,t})) - e(I'_t\,(s_{i,t}))\right\|^2 + \lambda_1(s_{i,t}-s_{i-1,t})^2 + \lambda_2(s_{i,t}-s_{i,t-1})^2, \quad (12)$$

where λ_1 and λ_2 are weights. Then we mask the pair of aligned video frames according to the calculated seam S_t^* and apply the multi-band blending algorithm [14] to synthesize the 360° video.

4. EXPERIMENTAL RESULTS

We test our video stitching method by using the video data captured by a Ricoh Theta S device, which has two fisheye cameras with 200° field of view. Three experiments are performed. The first one evaluates the feature alignment and the mesh deformation quality, the second one tests the performance of the proposed adaptive seam carving, and the third one evaluates the quality of the stitched videos. Our method is compared with two state-of-the-art stitching methods [7], [11] for the fisheye camera.

Fig. 4 compares the results of the first experiment obtained by the traditional method [7] and our method for six scenes. In Fig. 4(a), the accuracy is measured by the average error of the aligned features of the image pair. Our method has a lower average alignment error and better stitching quality. In Fig. 4(b), the quality is evaluated by measuring the displacement of the deformed mesh vertices between two consecutive frames. Our method has a smaller average vertex displacement and hence smoother 360° video output. It can be seen that our method is more robust than the traditional method in both feature alignment and mesh deformation.

Fig. 5 shows the seam generated by a previous method [11] and the proposed method for an example scene. It can be clearly seen that, unlike the previous method, our method smoothly cuts the flat planes (e.g. the curtain and the ceiling) in the scene and avoids generating discontinuities. This explains why our method is robust to noise and why the resulting video is jitter-free.

Fig. 6 shows the close-ups of three consecutive 360° video frames. It can be seen from the face and clothes in the image that our method generates seamless and stable stitching. The resulting video is free of ghost and fragmentation artifacts.

5. CONCLUSION

We have described a panoramic video stitching method for dual fisheye cameras. Our method preserves the temporal coherence by imposing a mesh deforming operation prior on the objective function. Unlike existing video stitching methods, our method is robust to object motion and noise. It optimizes the seam carving in both spatial and temporal dimensions. The 360° video generated by our method is seamless and free of jittering and flickering.

6. ACKNOWLEDGEMENTS

This work was funded in part by the Telecommunication Laboratories of Chunghwa Telecom Co. Ltd. under Contract TL-108-8301.

7. REFERENCES

[1] "RICOH THETA." [Online]. Available: https://theta360.com/en/. [Accessed: 28-Jan-2019].

[2] "Rylo 360° camera." [Online]. Available: https://www.rylo.com/. [Accessed: 28-Jan-2019].

[3] "Insta360 360 Camera." [Online]. Available: https://www.insta360.com/. [Accessed: 28-Jan-2019].

[4] F. Perazzi *et al.*, "Panoramic Video from Unstructured Camera Arrays," *Comput. Graph. Forum*, vol. 34, no. 2, pp. 57–68, May 2015.

[5] K. Lin, S. Liu, L. F. Cheong, and B. Zeng, "Seamless Video Stitching from Hand-held Camera Inputs," *Comput Graph Forum*, vol. 35, pp. 479–487, 2016.

[6] W. Jiang and J. Gu, "Video stitching with spatial-temporal content-preserving warping," in *2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2015, pp. 42–48.

[7] T. Ho, I. D. Schizas, K. R. Rao, and M. Budagavi, "360-degree video stitching for dual-fisheye lens cameras based on rigid moving least squares," in *2017 IEEE International Conference on Image Processing (ICIP)*, 2017, pp. 51–55.

[8] D. G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Nov. 2004.

[9] H. Bay, T. Tuytelaars, and L. Van Gool, "SURF: Speeded Up Robust Features," in *Computer Vision – ECCV 2006*, 2006, pp. 404–417.

[10] C. Liu, J. Yuen, A. Torralba, J. Sivic, and W. T. Freeman, "SIFT Flow: Dense Correspondence across Different Scenes," in *Computer Vision – ECCV 2008*, vol. 5304, D. Forsyth, P. Torr, and A. Zisserman, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 28–42.

[11] I.-C. Lo, K.-T. Shih, and H. H. Chen, "Image Stitching for Dual Fisheye Cameras," in *2018 25th IEEE International Conference on Image Processing (ICIP)*, 2018, pp. 3164–3168.

[12] S. Avidan and A. Shamir, "Seam Carving for Content-aware Image Resizing," *ACM Trans Graph*, vol. 26, no. 3, Jul. 2007.

[13] HSL, "A collection of Fortran codes for large scale scientific computation." [Online]. Available: http://www.hsl.rl.ac.uk/. [Accessed: 08-May-2019].

[14] M. Brown and D. G. Lowe, "Automatic Panoramic Image Stitching using Invariant Features," *Int. J. Comput. Vis.*, vol. 74, no. 1, pp. 59–73, Aug. 2007.