IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 1 January 2024, accepted 23 January 2024, date of publication 7 February 2024, date of current version 14 February 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3363882

RESEARCH ARTICLE

Stitching Videos From Unstructured Camera Arrays With Rectangular Boundaries

RUIFANG PAN[1,2], YUN ZHANG[2], LIN XU[1], AIHONG QIN[2], AND HUI DU[2]

[1]Information College, Zhejiang Guangsha Vocational and Technical University of Construction, Jinhua 322100, China
[2]College of Media Engineering, Communication University of Zhejiang, Hangzhou 310018, China

Corresponding author: Yun Zhang (zhangyun@cuz.edu.cn)

This work was supported in part by the Zhejiang Province Basic Public Welfare Research Program under Grant LGG22F020009, Grant LGF22F020015, and Grant LGF21F020002; and in part by the Key Laboratory of Film and TV Media Technology of Zhejiang Province under Grant 2020E10015.

ABSTRACT This paper presents a novel warping based method to stitch videos from unstructured camera arrays. Our approach adopts a two-step energy optimization for video stitching. In the first step, we perform an initial stitching on keyframes, and then extract the boundary vertices and warped vertices as constraints for further optimization. In the second step, we design a global optimization to effectively propagate the stitching from the keyframe to other frames while ensuring the feature alignment, boundary regularity and temporal coherence. The optimization can be efficiently solved by a linear system, and the final stitching results are produced by warping and blending. Experimental results and comparisons show that our method can efficiently stitch multiple videos from unstructured camera arrays, and outperforms state-of-the-art methods.

INDEX TERMS Video stitching, unstructured camera arrays, boundary regularity, temporal coherence, optimization.

I. INTRODUCTION

With the rise of VR/AR, videos with large field of view have become more and more popular due to their immersive and interactive experiences. However, our consumer-level cameras, such as smart phones and digital cameras, usually have limited field of view, making the video viewing less immersive. In recent years, many professional cameras can shoot videos with extremely large field of view, e.g. Nokia OZO, Samsung Gear 360, GoPro, Vuze etc., which can capture 360-degree panoramic videos. Unfortunately, these professional devices are very expensive and far from being as popular as consumer-level cameras. Therefore, stitching multiple input videos together to obtain a wide-angle panoramic video is a good choice for ordinary users, and has been widely researched in recent years.

Stitching aims to solve the field of view (FOV) limitation of images/videos, and has been used in various fields such as sports broadcasting, video surveillance, street view [1]. Recently, many researches [2], [3], [4] focused on stitching videos captured by several freely hand-held cameras, which are challenged by the shakiness of captured videos, large parallax between videos, complex foreground and background etc. Although successful in many freely captured videos, video stitching still suffers from many drawbacks: (1) the combined stabilization and stitching are very difficult to optimize, and the process is extremely time-consuming; (2) it is hard to collaborate the shooting process of each hand-held camera; (3) the final stitched video usually has very jittery irregular boundaries, which may greatly reduce the video contents after cropping.

Compared with the freely hand-held video capture, camera arrays that fix multiple cameras on rigs, see Fig. 1, can better collaborate the shooting of multiple cameras, and largely reduce the difficulty of stitching. Perazzi et al. [5] proposed a method to generate panoramic videos from unstructured camera arrays. Lai et al. [6] further proposed a wide baseline video stitching algorithm for linear camera arrays. Although seamless and visually pleasing, the stitched videos always have irregular boundaries, due to the unstructured camera

The associate editor coordinating the review of this manuscript and approving it for publication was Chaker Larabi.

arrays, and there might be much video content loss after cropping. To preserve the video content after the geometric warping, Zhang et al. [7] formulated image stitching and boundary regulation in a unified optimization framework, and further applied it to videos captured by unstructured camera arrays. However, they used the same parameters across neighboring frames, which cannot ensure accurate feature alignment. Inspired by [8], Wu et al. [9] recently proposed a warping-based approach for rectangling irregular videos. Although effective in many examples, their method suffers from the seam insertion in video mesh placement, which is extremely time-consuming. In addition, the video meshes may contain regions outside the stitched video frames, leading to "holes" in result videos after rectangling.

In this paper, we propose a temporal-spatial coherent warping to stitch videos from unstructured video arrays. Unlike videos from freely hand-held cameras, videos captured from fixed camera arrays are much easier to stitch, because the relative position between cameras remains unchanged. Our key observation is that continuous video frames undergo similar mesh transformation in stitching and rectangling. Thus, we first perform the stitching on a keyframe (e.g. the first frame) using the method in [7], and the warped vertices will be used as a reference for other frames. Then, we construct a global energy function for video stitching, with mesh propagation, feature alignment, rectangular boundary as constraints, and the warped meshes are obtained by the energy optimization. Finally, the video stitching result is obtained by texture mapping and video blending.

Compared with previous stitching and rectangling methods [7], [9], our method can reach a good balance between efficiency and visual effects, and the main contributions can be listed as follows:

- We propose the first warping-based optimization to stitch videos captured by unstructured camera arrays, which can produce temporally coherent and visually pleasing video stitching results with rectangular boundaries.
- For efficient and content-preserving video stitching, we design a two-step optimization scheme. We first perform stitching and rectangling on a keyframe, then propagate the mesh on the keyframe to neighboring frames by designing a global optimization with feature alignment, mesh propagation, rectangular boundary as constraints.

The organization of this article is as follows. We first briefly review the techniques related to our work in Section II. Then we describe the detailed algorithm of video stitching and rectangling as well as the implementation details in Section III, and evaluate the performance of our method in Section IV. Finally, we conclude the paper in Section V.

II. RELATED WORK

Image and video stitching has been extensively researched in the field of computer graphics and computer vision to solve the problem of limited field of view in images and videos [1]. In this section, we briefly review the techniques most related to our work.

A. IMAGE STITCHING

Image stitching refers aligning, blending multiple images with overlapped regions to generate a new image with wide field of view, which has been widely researched for decades, and successfully applied to many portable devices, such as smart phones, digital cameras. In general, image stitching approaches can be divided into two main categories: warping-based and seam-driven methods. In the warping-based methods, multiple models are usually used to represent the corresponding relationship between images, and the feature alignment, local and global similarity are achieved by the grid warping guided by the global and local optimization. Lin et al. [10] proposed a smoothly varying affine transformation for locally adaptive image stitching, which also preserves the global similarity. For better alignment, Zaragoza et al. [11] proposed an As-Projective-As-Possible warping to adjust local regions that are inconsistent with the global projective model. To reduce the distortions introduced by the warping-based stitching, Chen et al. [12] proposed a local warping model with the global similarity as constraint, which makes stitching results more natural. Li et al. [13] proposed robust elastic warping, which can tolerate parallax in image stitching. The seam-driven method aims to find an optimal seam in the roughly aligned regions, so as to deal with the large parallax in stitching. Zhang et al. [14] believed that the overlapped regions do not need to be precisely aligned. They proposed a homography and content-preserving warping to deal with large parallax, and further put forward a method to find an optimal seam in the overlapping regions. In view of the shortcomings of previous seam estimation after the feature alignment, Lin et al. [15] proposed a seam-estimation to guid the optimization in local feature alignment, and improved the stitching result iteratively. To diminish the large parallax, Xue et al. [16] proposed a stable hybrid actor-critic to estimate stable seam measurements in the overlapping region. Recently, the deep learning framework has been applied to image stitching. Zhao et al. [17] proposed a deep neural network to accurately estimate the homography of image stitching with small parallax, and a new stitching loss function for content preserving. To deal with the limitations of few features and lack of labeled data, Nie et al. [18] proposed an unsupervised deep image stitching framework, which can generate comparable results to supervised methods.

B. VIDEO STITCHING

According to the configuration of multiple cameras, video stitching methods can be divided into two types: relatively stationary and freely moving. When cameras are relatively stationary, videos are globally stitched by pre-calibrating the relative positions of cameras. Li et al. [19] proposed

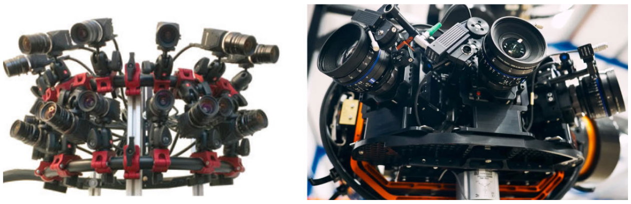


FIGURE 1. **Two types of camera arrays. Left: unstructured camera arrays constructed in [5]; Right: structured camera arrays to capture 12K resolution videos, refer to https://ymcinema.com/2022/01/31/worlds-first-red-v-raptor-8k-aerial-camera-array/.**

an efficient video stitching method by using fast structure deformation. Jiang et al. [20] proposed a spatial-temporal content-preserving warping to stitch multiple synchronized videos. Perazzi et al. [5] proposed a weighted extrapolation of warps for panorama videos stitching from unstructured camera arrays. Compared with video stitching from relatively stationary cameras, stitching videos from freely moving cameras is more complex due to the Intra and inter frame motions. Lin et al. [4] proposed to stitch videos captured by hand-held cameras by warping, which is achieved by optimizing the temporal stability and alignment quality. To reduce the spatial and temporal artifacts when stitching the shaking videos, Guo et al. [3] proposed a unified framework to jointly perform video stitching and stabilization. Nie et al. [2] improved the unified stitching and stabilization by identifying the background and eliminating the false feature matches.

C. RECTANGLING THE STITCHING RESULTS

Rectangling aims to generate rectangular images from the stitching results with irregular boundaries. He et al. [8] are the first to propose a warping based rectangling to generate rectangular images from the stitched images with irregular boundaries. Inspired by [8], Wu et al. [9] further proposed a spatio-temporal warping based rectangling to rectify the videos with irrecgular boundaries. Different from [8] and [9], which take the stitched results as input, Zhang et al. [7] combined image stitching and rectangling in a global optimization, which can produce panoramic images with natural alignment and regular boundaries. Recently, Nie et al. [21] proposed the first deep learning solution to image rectangling, which can produce rectangular images in a residual manner, and preserve linear and non-linear structures. Following [21], Nie et al. [22] further proposed a learning-based method to correct rotation in images without angle prior, which can automatically correct tilted images by regressing the mesh deformation.

III. ALGORITHM

Previous video stitching methods can provide spatial aligned and temporal coherent results [2], [3], however, their methods are too complicated and extremely time-consuming and cannot preserve rectangular boundaries, which limits their practical use. In this paper, the videos to be stitched are shot by unstructured cameras, and we assume that there is unnoticeable shakes and relative movement in cameras, thus the stitched video frames usually have fixed boundaries. The naive extension of image stitching to video frames may introduce discontinuities between frames, see Fig.5. In this paper, we propose a novel and effective solution for efficient stitching. Fig. 2 gives the pipeline of our video stitching method. The input to our method is a number of videos with partial content overlaps, and the goal is to obtain a panoramic video with rectangular boundaries. Like previous warping-based stitching, we also place a quad mesh on each frame, and the stitching result is obtained by warping the meshes guided by the constraints on them. We first divide video frames into several blocks with overlaps, and select a keyframe in each block (the first frame by default), and perform stitching and rectangling using an energy optimization on quad meshes. After obtaining the warped mesh on keyframes, we further propagate them to other frames of the block while ensuring the feature matching and boundary regularity, and the final mesh of each frame can be efficiently calculated by the energy optimization in each frame.

A. KEYFRAME STITCHING AND RECTANGLING

Fig. 3 shows the flowchart of keyframe stitching and rectangling. Inspired by [7], the keyframe is first initially stitched using traditional mesh-warping based optimization method [12], which aims to obtain the warped mesh of each stitched image. Then, the irregular boundary of the stitching result is obtained by the polygon Boolean union operation [23] of each warped mesh, and we further construct the rectangular boundary constraint based on this. Finally, we construct a global optimization with feature matching, shape preserving, rectangular boundary preserving as constraints, and the final warped mesh is obtained by solving a linear system whose number of unknowns is proportional to the number of mesh vertices.

1) INITIAL STITCHING

Similar to many warping-based methods [2], [3], [12], we place a quad mesh on each image to be stitched. Let $V = \{V^i\}$ and $E = \{E^i\}$ be the set of vertices and edges of the keyframes captured from different cameras, where $i = 1, 2, \ldots, N$, and N is the number of cameras. We aim to obtain the warped vertices of the quad mesh by minimizing the energy functions with feature alignment, local and global similarity as constraints. Inspired by [7] and [12], we define each energy term as follows.

2) FEATURE ALIGNMENT

Traditional methods may fail to match features in textureless or ambiguous regions, and they also consume much time and memory for the complex calculation. For robust and accurate feature alignment, we use state-of-the-art learning-based method [24] for feature matching between images, and

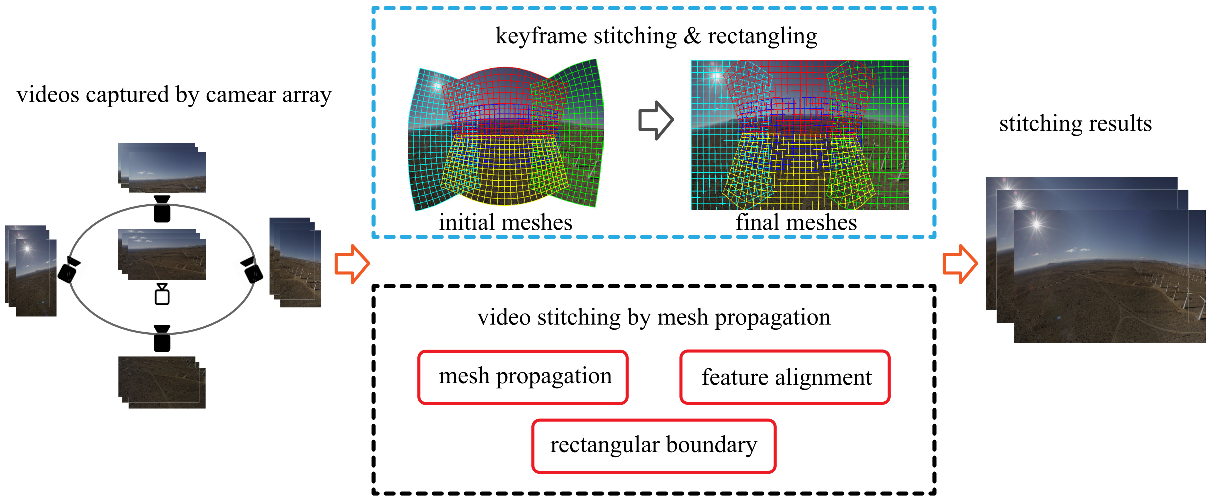

videos captured by camear array
keyframe stitching & rectangling
initial meshes
final meshes
stitching results
video stitching by mesh propagation
mesh propagation
feature alignment
rectangular boundary

FIGURE 2. Flowchart of our method. The input to our method is a number of videos with partial overlaps, and the goal is to obtain a panoramic video with a rectangular boundary. We first divide video frames into several blocks with overlaps, and select a keyframe in each block (the first frame by default), and perform stitching and rectangling using the energy optimization on quad meshes. After obtaining the warped meshes on the keyframe, we further propagate them to other frames of the block while ensuring the feature alignment and rectangular boundary constraints.

the energy term is defined as

$$\zeta_{align}(V) = \sum_{\{i,j\}} \sum_{(\delta_k^i, \delta_k^j) \in \Delta^{i,j}} ||\Theta(\delta_k^i) - \Theta(\delta_k^j)||^2, \qquad (1)$$

where $\{i, j\}$ refers to all matching image pairs with overlaps. $\Delta^{i,j}$ enumerates all matched features between image i and j. To constrain the feature alignment between images on the quad mesh, we represent each matched feature using the interpolation of the vertices of the mesh grid that contains the feature point. E.g. $\Theta(\delta_k^i)$ refers to the bilinear combination of the vertices of the mesh grid that contains the feature δ_k^i.

a: LOCAL AND GLOBAL SIMILARITY

In addition to feature alignment, we also keep the local and global similarity to reduce unwanted distortions and make the stitching as natural as possible. Local similarity aims to preserve the shape of the quad mesh. Similar to [25], we split each quad mesh into 2 triangles, and constrain the shape of triangles, which can be easily implemented as:

$$\zeta_{loc\_sim}(V) = \sum_{i=1}^{N} \sum_{V_k^i} \|V_k^i - V_{k_1}^i - \eta\Psi(V_{k_0}^i - V_{k_1}^i)\|^2, \quad (2)$$

where the scaling factor $\eta = \|V_k^i - V_{k_1}^i\| / \|V_{k_0}^i - V_{k_1}^i\|$ and Ψ is a 90° rotation matrix. To make the stitching as natural as possible, global similarity is used to optimize the rotation and scaling factors of each image in stitching. We use the energy term defined in [12], and the desired scaling s_i and rotation angle ϑ are calculated w.r.t. the reference image (normally the first image), and the energy term is defined as:

$$\zeta_{gl\_sim}(V) = \sum_{i=2}^{N} \sum_{e_j^i \in E^i} \gamma(e_j^i)(\|w_x(e_j^i) - s_i \cos(\vartheta_i)\|^2 + \|w_y(e_j^i) - s_i \sin(\vartheta_i)\|^2), \qquad (3)$$

where $w_x(e_j^i)$ and $w_y(e_j^i)$ are the weights of grid edges to ensure a similarity transform in x and y directions, and $\gamma(\cdot)$ is used to emphasize the edges in overlapping regions.

Finally, the energy functions can be defined by simply combining the feature alignment and local&global energy terms above in a linear weighting manner:

$$S_{init}(V) = \beta_a \zeta_{align}(V) + \beta_l \zeta_{loc\_sim}(V) + \beta_g \zeta_{gl\_sim}(V), \quad (4)$$

where β_a, β_l, β_g are used to balance the importance of each energy term.

b: KEYFRAME RECTANGLING

After the initial stitching step, images are well stitched but always have irregular boundaries. To obtain the stitching results with rectangular boundaries, we have to further consider the boundary constraint in stitching. Similar to [7], we first obtain the outer boundaries of each mesh, and take them as a polygon $\hat{\Omega}^i$; then we obtain the outer boundary vertices Ω of the stitching result by polygon union operators:

$$\Omega = \bigcup_{i=1}^{N} \hat{\Omega}^i. \qquad (5)$$

Finally, we set the bounding rectangle $Q(\hat{\Omega}^i)$ for the outer boundary, and select 4 vertices $\{\Gamma_p\}, p \in \{1, 2, 3, 4\}$ from Ω that are closest to the 4 corners of $Q(\hat{\Omega}^i)$.

With the outer boundary vertices $\hat{\Omega}$ and the four corners $\{\Gamma_p\}$, we can easily classify the vertices on the outer boundary into four different sides (top, bottom, left, right), and the rectangling is achieved by dragging the vertices on each side to the corresponding outer rectangle $Q(\hat{\Omega}^i)$. We record the target value and direction of the outer boundary vertices Ω as $\varepsilon(\Omega)$ and $D(\Omega)$.

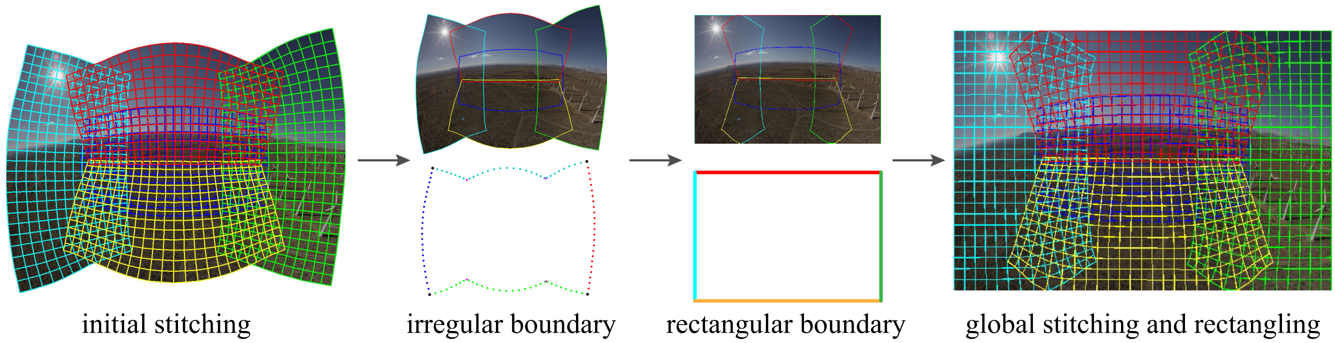


FIGURE 3. Keyframe stitching and rectangling flowchart. The images of the keyframe are first stitched by traditional image stitching method, which produces irregular boundaries, then the irregular boundary is extracted by the polygon Boolean union operator, and the rectangling is achieved by enforcing the rectangular boundary constraint in the energy optimization.

Similar to [7], the rectangular boundary constraint is defined as:

$$\zeta_b(V) = \sum_{k=1}^{M} \|(\Omega_k \cdot D(\Omega_k)) - \varepsilon(\Omega_k)\|^2, \tag{6}$$

where M is the number of vertices on the outer boundary, and $D(\Omega_k)$ is used to project Ω_k on x or y directions by setting it to $[1, 0]$ or $[0, 1]$ respectively, and $\varepsilon(\Omega_k)$ records the target values of each vertex on left, top, right, or bottom directions.

B. TEMPORAL COHERENT VIDEO STITCHING BY MESH PROPAGATION

In [7], video stitching with rectangular boundaries is simply achieved by appling the same parameters to a set of continuous frames, which does not consider the temporal coherence in video stitching. To keep the motion coherence in neighboring frames, a direct idea is to track features in consecutive frames, however, it is time-consuming to track feature trajectories in several video frames, and the constraints would be too complex to be optimized. In this paper, we propose a simple and effective method to ensure the temporal coherence of video stitching results. For the keyframe of each block, we further perform rectangling based on the initial stitching result using the SOTA method [7], and obtain the final warped mesh vertices $\{\tilde{V}^i\}$. With the stitching and rectangling result in the keyframe of each block, we further propagate their mesh vertices to the following frames of this block, by enforcing that the mesh vertices of each frame are close to that of the keyframe, and define it as follows:

$$\zeta_p(V) = \sum_{i=1}^{N} \sum_{j} \|V_j^i - \tilde{V}_j^i\|^2. \tag{7}$$

Finally, the meshes of video stitching result can be obtained by optimizing the following energy function:

$$S_{final}(V) = \sum_{t} (\beta_a \cdot \zeta_a + \beta_b \cdot \zeta_b + \beta_p \cdot \zeta_p), \tag{8}$$

where β_b, β_p are the weights to specify the importance of the rectangular boundary and temporal coherence terms. $\{\tilde{V}^i\}$ is updated after stitching each frame. t enumerates all frames in a block. In this step, the local and global shape preserving energies are not required due the use of the temporal coherence term, which not only simplifies the optimization but ensures an robust and easy-to-control stitching.

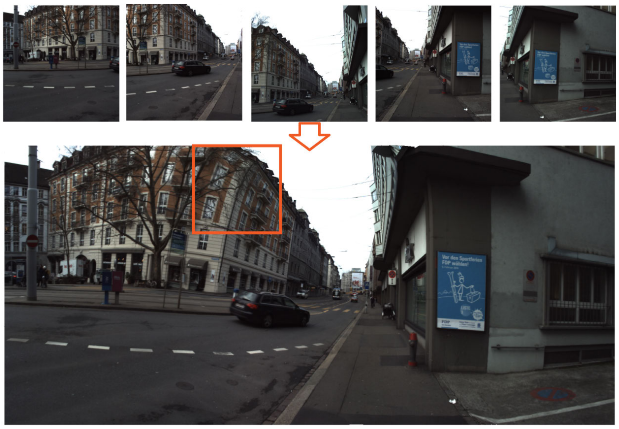


FIGURE 4. Failure case. Our method may fail when there is large content missing or salient structure near the boundary.

C. IMPLEMENTATION DETAILS

In the initial stitching step, we set $\beta_a = 1$, $\beta_l = 0.75$, $\beta_g = 20$ and all examples work well. In the temporal coherent stitching step, we set $\beta_a = 1$, $\beta_b = 20$, $\beta_p = 5$ for all frames. For more robust and coherent results, we split video frames into several blocks with overlaps in the temporal dimension, and each block contains 30 to 40 frames. We select the first frame of each block as the keyframe, and the keyframe images of different views are stitched using SOTA stitching method [7]. To obtain temporal coherent results, we further linearly interpolate the warped mesh of frames in the overlapping area of adjacent blocks, and the final stitching results are produced by mesh based re-rendering and image blending.

IV. EXPERIMENTS AND EVALUATIONS

In this section, we show video stitching results of our method and comparisons with SOTA methods. Then,

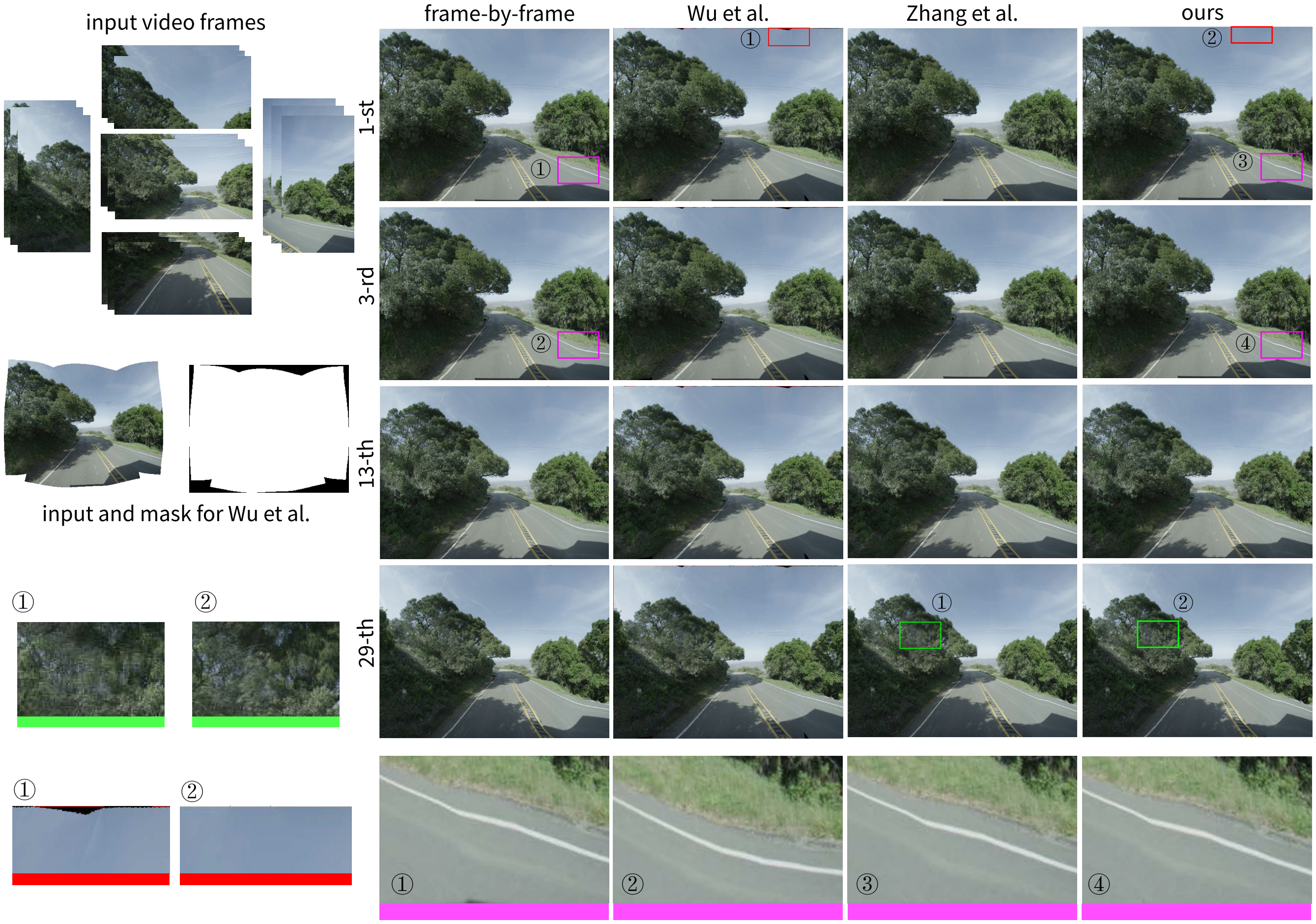


FIGURE 5. Results and comparisons with SOTA methods. The left column shows input video frames, the right columns present stitching result of different frames by different methods. We also provide zoom-in views for detailed comparisons, and use indices and colors to identify different boxes.

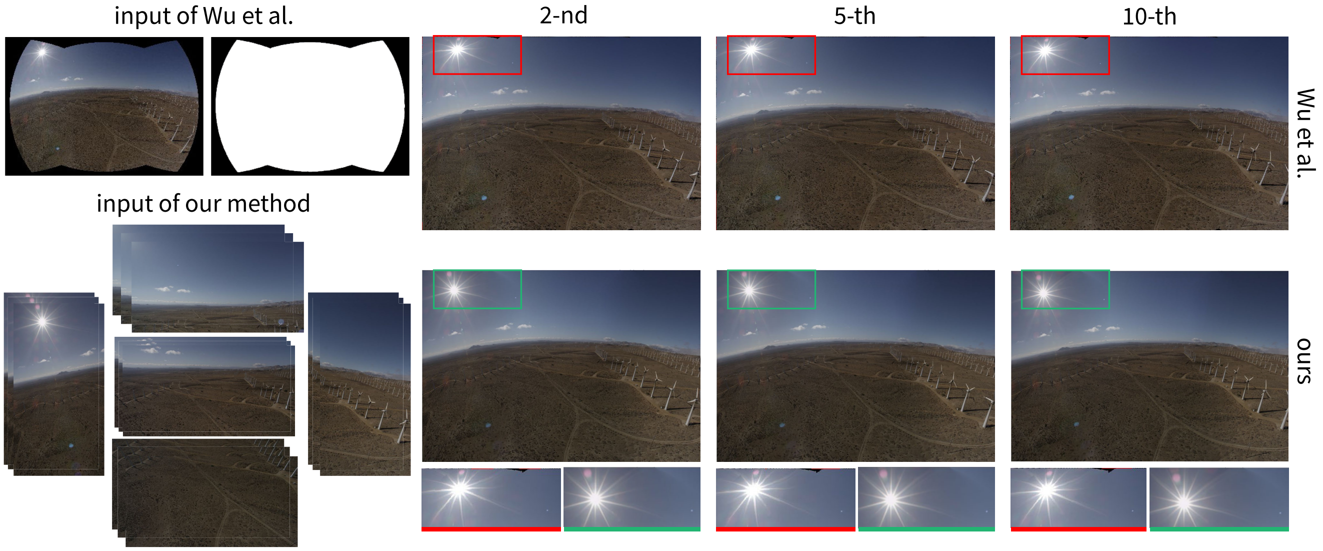


FIGURE 6. Comparison with Wu et al's method [9]. The left column shows the input frames our method, and the initial stitched frames and masks of Wu et al's method [9]. The right 3 columns give stitching results of different frames by different methods, and the zoom-in views further provide the comparisons.

we further report the performance and qualitative evaluations, and ablation study to show the effectiveness of our method. We make use of the video data from [5] for all experiments in this paper, and the comparisons

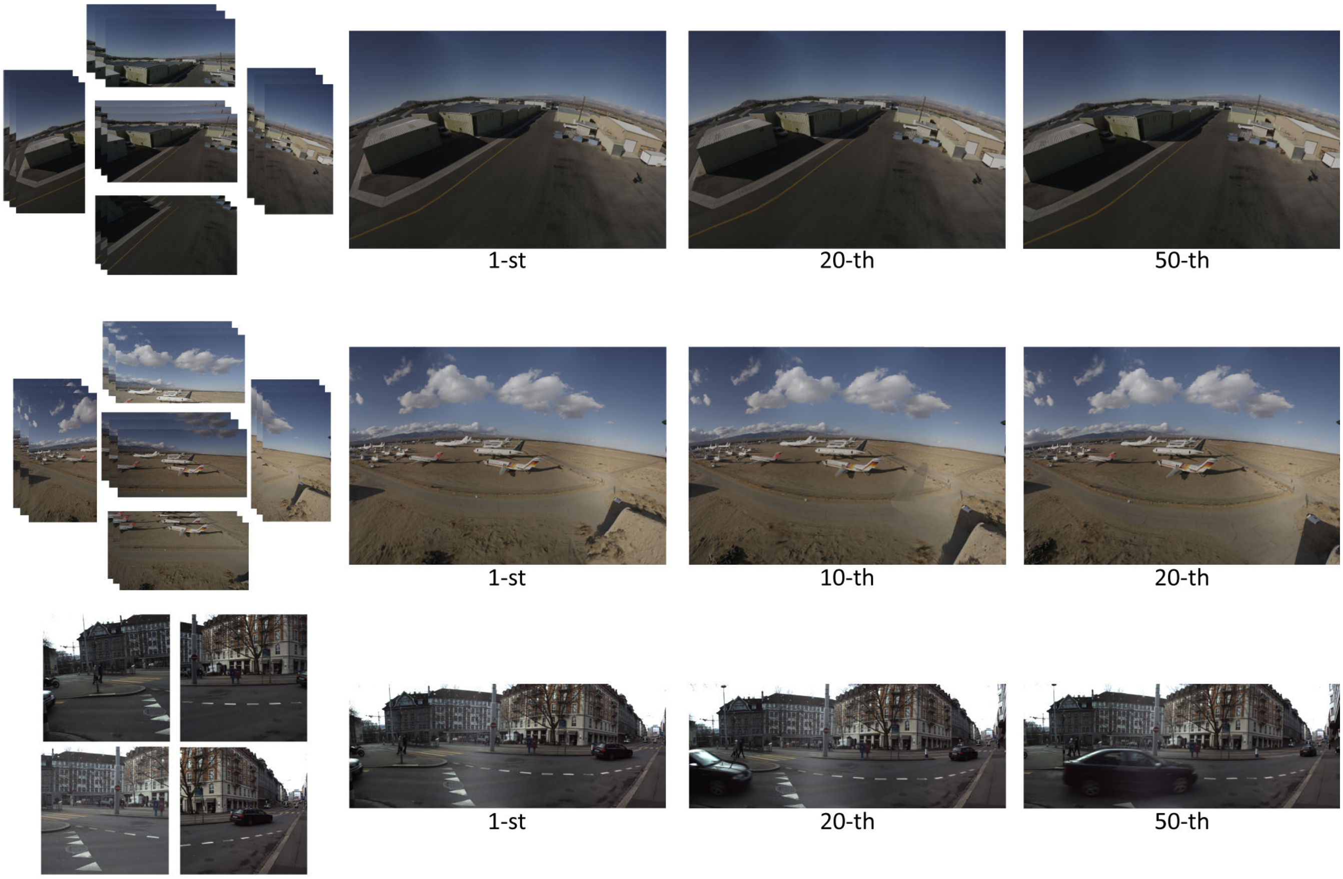


FIGURE 7. More results. The 1^{st} column shows the source videos of different views, and the 2^{nd} to 4^{th} columns present the stitching results of different frames.

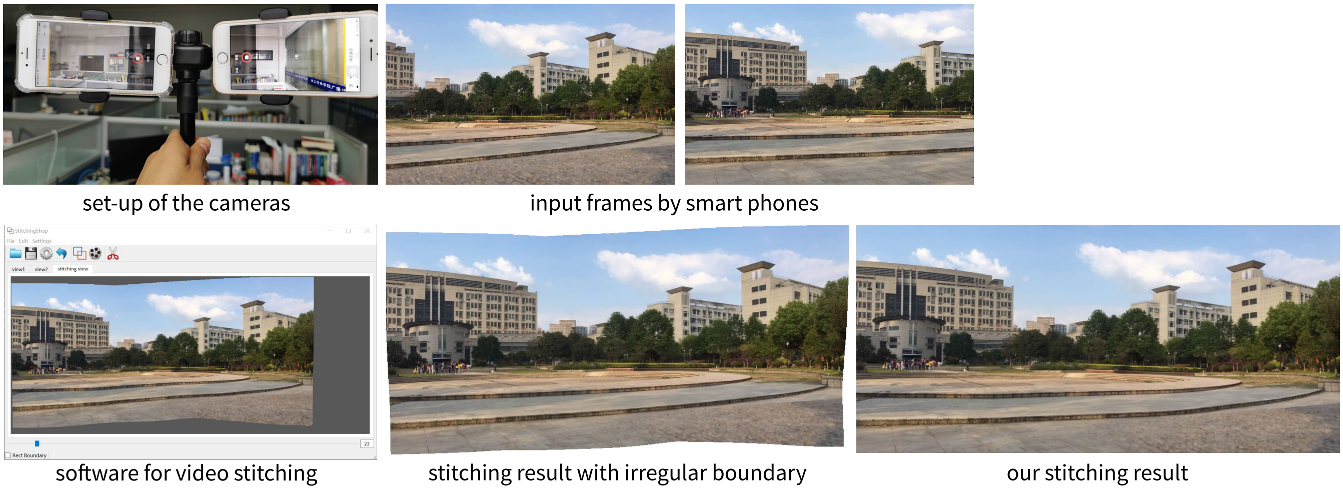

set-up of the cameras
input frames by smart phones
software for video stitching
stitching result with irregular boundary
our stitching result
stitchingView
File Edit Settings
view1 view2 stitching view
Rect Boundary

FIGURE 8. Stitching videos shot by smart phones. The two smart phones are fixed on a bracket, and videos are shot simultaneously through the two smart phones. We also developed a software for the video stitching, and gave stitching result with irregular and rectangular boundary.

are produced by the source code provided by their authors.

A. RESULTS AND COMPARISONS

Fig. 5 shows video stitching results by our method and comparison with frame-by-frame stitching, Zhang et al.'s method [7] and Wu et al.'s method [9]. The input video frames are from 5 cameras, which are fixed as unstructured camera arrays. For Wu et al.'s method [9], the input video are initially stitching results by traditional video stitching method and the corresponding masks. For Zhang et al.'s method [7], video frames are stitched by the warping parameters of keyframes. We perform stitching using 4 different methods, and the results and zoom-in views vividly show the advantages of our

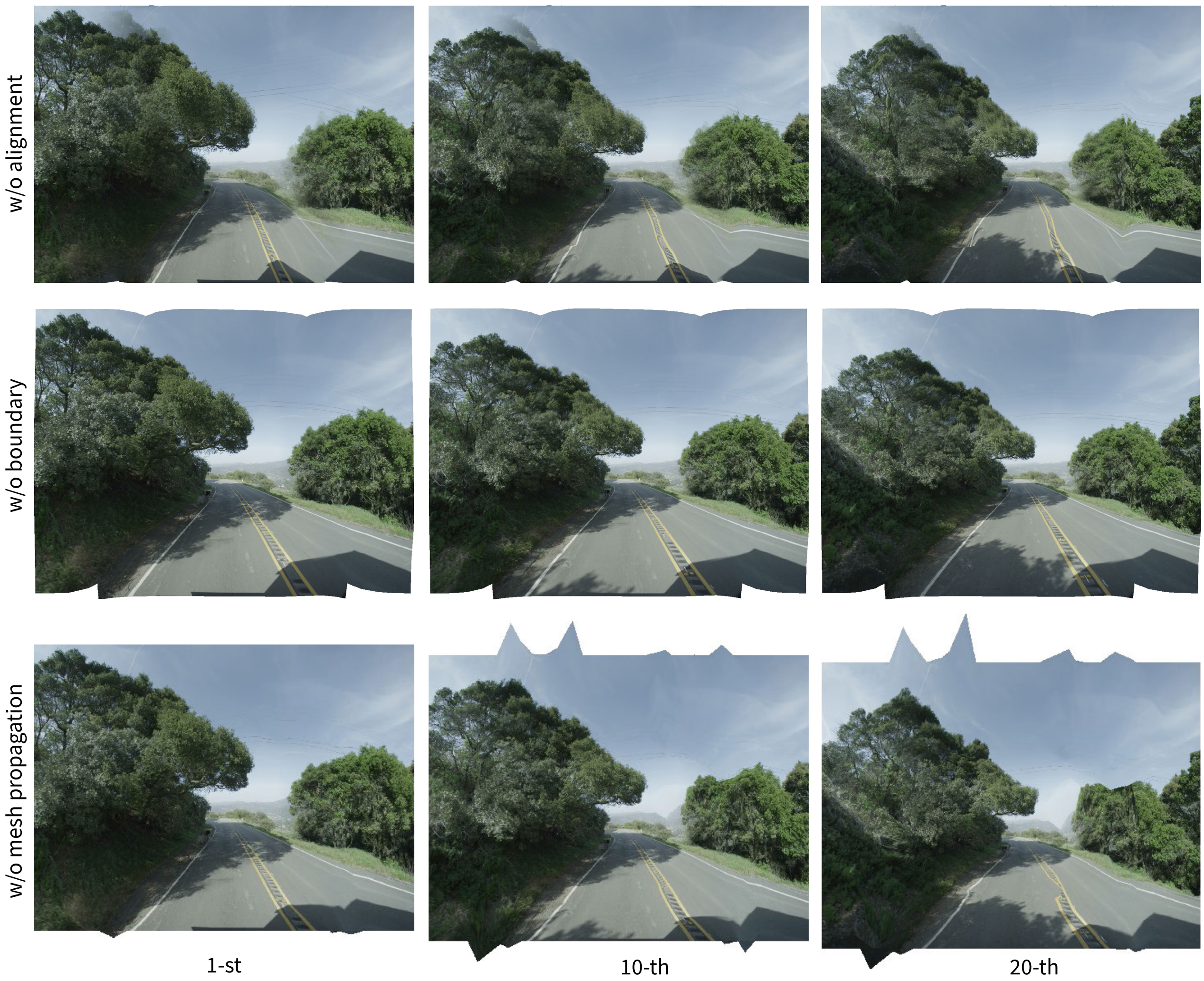


FIGURE 9. **Ablation study. To show the effects of each energy term, we give stitching results of different frames without the feature alignment, rectangular boundary and mesh propagation constraints, respectively.**

method. The frame-by-frame method may produce temporal discontinuous results; Wu et al.'s method [9] cannot ensure rectangular boundaries; Zhang et al.'s method [7] can not ensure good feature alignment due to the parameter sharing in continuous frames.

In Fig. 6, we further compare our method with Wu et al. [9]. Comparisons and zoom-in views in the red and green boxes show that our method outperforms [9] in terms of structure preserving and regular boundary preserving. More video stitching results provided in Fig. 7 show that our method can produce high-quality video stitching results while preserving the rectangular boundaries, and well adapt to different kinds of scenes.

To stitch videos shot by smart phones, we designed a special camera set-up, which can fix several smart phones on a bracket, see Fig. 8. In this camera set-up, videos of different views can be shot simultaneously by the well-designed apps on each cellphone. We also developed a software system to stitch videos from smart phones, and stitch videos with irregular and rectangular boundaries. Comparisons show that our stitching with rectangular boundary has better wide-angle and visual effects.

B. ABLATION STUDY

We conduct an ablation study to test the effects of each energy term for video stitching. As shown in Fig. 9, without the alignment term, video are not correctly stitched in the overlapping regions; without the rectangular boundary term, we cannot expect the stitching results with rectangular boundaries; without the mesh propagation term, the stitching results cannot well preserve the shape.

C. PERFORMANCE

We report the performance of our method on an Intel Core i7 12700H 2.3GHz laptop with 32G RAM. Take the experimental results in Fig. 5 as an example, which contains input videos from 5 different cameras, and each video has 30 frames with a resolution of 800 × 600. The total time cost is

TABLE 1. Comparison of running time (Sec.). We compare our method with the frame-by-frame, Zhang et al. [7] and Wu et al. [9] using the examples in Figs. 5 and 6. We give the running time for the examples by each method. For Wu et al.'s [9] method, their running time consists of the seam searching and video warping.

	frame-by-frame	Zhang et al. [7]	Wu et al. [9]	ours
Fig.5	188.8s	45.1s	2550s+288s	181.1s
Fig.6	161.56s	42.6s	1983s+247s	142.6s
Fig.7-1	153.1s	51.6s	2180s+235s	146.2s
Fig.7-2	162.2s	31.5s	2360s+225s	135.7s
Fig.7-3	135.4s	41.9s	2410s+245s	120.1s

181.1 sec. We give the comparison of running time in Table 1, which shows the performance of examples in this paper. To make a fair comparison, the number of pixels of each image frame is normalized to be 800 × 600, and the number of frames is 30 for all examples. The first column shows the performance of results by frame-by-frame stitching, and the second and third column give performance of stitching by [7] and [9]. From comparison we find that the method of [7] is the most efficient due to the stitching parameter sharing in neighboring frames. The most time-consuming method is [9], and the running time consists of two parts: seam searching and temporal consistent video warping, in which seam searching consumes a significant amount of time. Compared with the frame-by-frame method, our method is more efficient, because our energy function considers the temporal coherence by mesh propagation that makes it easier to optimize.

D. USER STUDY

To evaluate the quality of video stitching results, we invited 30 students from our university aging from 20 to 23, and asked them to give scores for stitching results by Perazzi et al. [5], and rectangular stitching (frame-by-frame, Wu et al. [9] and our method) in terms of wide-angle effects and visual effects. In order to give scores more objectively and accurately, we first told them the main indicators for visual effect evaluation of video stitching, which include distortions, structure preserving, temporal coherence. Then, we gave the cropping ratios of each stitching result, which is the average ratio of the cropped content by a rectangle to the whole stitched panorama. With the cropping ratios as reference, they can evaluate the wide-angle effects more easily and accurately. In our user study, each indicator is given by an integer ranging from 0 to 5 (worst to best). We give average scores of all participants for all examples in Table 2. In terms of the wide angle effects, the method in [5] has the worst performance, due to not considering the regularity of boundaries, and thus less content are preserved after being cropped by a rectangle. For the visual effects, the frame-by-frame method fails to make satisfied results due to the failure to preserve the temporal coherence. The statistical data in the user study indicates the advantages of our method over the SOTA.

TABLE 2. User study of different methods. We give average scores of all participants for each examples by different methods, and each score includes wide-angle effects and visual effects.

	Perazzi et al. [5]	frame-by-frame	Wu et al. [9]	ours
Fig.5	3.54/4.01	4.22/3.89	4.36/4.46	4.71/4.68
Fig.6	3.21/4.12	4.36/4.23	4.51/4.49	4.81/4.72
Fig.7-1	3.75/4.35	4.33/4.25	4.41/4.50	4.49/4.55
Fig.7-2	3.68/4.11	4.13/4.33	4.42/4.51	4.52/4.63
Fig.7-3	4.01/4.11	4.21/4.16	4.22/4.51	4.75/4.81

V. CONCLUSION

This paper proposes a novel video stitching method to stitch videos taken from several unstructured camera arrays. We first divide videos into several blocks with temporal overlaps. Then, we perform initial stitching for the keyframe of each block, and further rectangling them using SOTA method. We further construct an energy function by enforcing feature alignment, rectangular boundary, and temporal coherence constraints, and obtain the optimized meshes for high-quality video stitching. Finally, the video stitching result is produced by mesh-based re-rendering and image blending. Different from previous video stitching methods [2], [9], which consume much computation and memory to stitch several frames together, our method only optimizes a single frame in the energy function, which is more efficient and easy to control. Experiments and comparisons on several examples show that our method is advantageous over existing SOTA methods in terms of speed and visual effects.

Our method still suffers from some limitations: (1) it is limited to stitching videos captures from unstructured camera arrays; (2) Our method cannot avoid large distortions near the stitching boundaries, and the distortion is severe when there is large content missing, see Fig.4.

In the future, we will turn to study the data-driven video stitching in the learning based framework, and a large dataset is required, and the network should be carefully designed which considers constraints, like feature alignment, structure preserving and temporal coherence.

REFERENCES

[1] W. Lyu, Z. Zhou, L. Chen, and Y. Zhou, "A survey on image and video stitching," *Virtual Reality Intell. Hardw.*, vol. 1, no. 1, pp. 55–83, 2019.

[2] Y. Nie, T. Su, Z. Zhang, H. Sun, and G. Li, "Dynamic video stitching via shakiness removing," *IEEE Trans. Image Process.*, vol. 27, no. 1, pp. 164–178, Jan. 2018.

[3] H. Guo, S. Liu, T. He, S. Zhu, B. Zeng, and M. Gabbouj, "Joint video stitching and stabilization from moving cameras," *IEEE Trans. Image Process.*, vol. 25, no. 11, pp. 5491–5503, Nov. 2016.

[4] K. Lin, S. Liu, L. Cheong, and B. Zeng, "Seamless video stitching from hand-held camera inputs," *Comput. Graph. Forum*, vol. 35, no. 2, pp. 479–487, May 2016.

[5] F. Perazzi, A. Sorkine-Hornung, H. Zimmer, P. Kaufmann, O. Wang, S. Watson, and M. Gross, "Panoramic video from unstructured camera arrays," *Comput. Graph. Forum*, vol. 34, no. 2, pp. 57–68, May 2015.

[6] W. Lai, O. Gallo, J. Gu, D. Sun, M. Yang, and J. Kautz, "Video stitching for linear camera arrays," in *Proc. 30th Brit. Mach. Vis. Conf. (BMVC)*, Cardiff, U.K.: BMVA Press, 2019, p. 130.

[7] Y. Zhang, Y. Lai, and F. Zhang, "Content-preserving image stitching with piecewise rectangular boundary constraints," *IEEE Trans. Vis. Comput. Graph.*, vol. 27, no. 7, pp. 3198–3212, Jul. 2021.

[8] K. He, H. Chang, and J. Sun, "Rectangling panoramic images via warping," *ACM Trans. Graph.*, vol. 32, no. 4, p. 79, 2013.

[9] J.-L. Wu, J.-J. Shi, and L. Zhang, "Rectangling irregular videos by optimal spatio-temporal warping," *Comput. Vis. Media*, vol. 8, no. 1, pp. 93–103, Mar. 2022.

[10] W. Lin, S. Liu, Y. Matsushita, T. Ng, and L. F. Cheong, "Smoothly varying affine stitching," in *Proc. 24th IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Colorado Springs, CO, USA, Jun. 2011, pp. 345–352.

[11] J. Zaragoza, T.-J. Chin, Q.-H. Tran, M. S. Brown, and D. Suter, "As-projective-as-possible image stitching with moving DLT," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 7, pp. 1285–1298, Jul. 2014.

[12] Y. Chen and Y. Chuang, "Natural image stitching with the global similarity prior," in *Proc. 14th Eur. Conf. Comput. Vis.* (Lecture Notes in Computer Science), vol. 9909, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Amsterdam, The Netherlands: Springer, 2016, pp. 186–201.

[13] J. Li, Z. Wang, S. Lai, Y. Zhai, and M. Zhang, "Parallax-tolerant image stitching based on robust elastic warping," *IEEE Trans. Multimedia*, vol. 20, no. 7, pp. 1672–1687, Jul. 2018.

[14] F. Zhang and F. Liu, "Parallax-tolerant image stitching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Columbus, OH, USA, Jun. 2014, pp. 3262–3269.

[15] K. Lin, N. Jiang, L. Cheong, M. N. Do, and J. Lu, "SEAGULL: Seam-guided local alignment for parallax-tolerant image stitching," in *Proc. 14th Eur. Conf. Comput. Vis.* (Lecture Notes in Computer Science), vol. 9907, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Amsterdam, The Netherlands: Springer, 2016, pp. 370–385.

[16] W. Xue, W. Xie, Y. Zhang, and S. Chen, "Stable linear structures and seam measurements for parallax image stitching," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 1, pp. 253–261, Jan. 2022.

[17] Q. Zhao, Y. Ma, C. Zhu, C. Yao, B. Feng, and F. Dai, "Image stitching via deep homography estimation," *Neurocomputing*, vol. 450, pp. 219–229, Aug. 2021.

[18] L. Nie, C. Lin, K. Liao, S. Liu, and Y. Zhao, "Unsupervised deep image stitching: Reconstructing stitched features to images," *IEEE Trans. Image Process.*, vol. 30, pp. 6184–6197, 2021.

[19] J. Li, W. Xu, J. Zhang, M. Zhang, Z. Wang, and X. Li, "Efficient video stitching based on fast structure deformation," *IEEE Trans. Cybern.*, vol. 45, no. 12, pp. 2707–2719, Dec. 2015.

[20] W. Jiang and J. Gu, "Video stitching with spatial–temporal content-preserving warping," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Boston, MA, USA, Jun. 2015, pp. 42–48.

[21] L. Nie, C. Lin, K. Liao, S. Liu, and Y. Zhao, "Deep rectangling for image stitching: A learning baseline," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New Orleans, LA, USA, Jun. 2022, pp. 5730–5738.

[22] L. Nie, C. Lin, K. Liao, S. Liu, and Y. Zhao, "Deep rotation correction without angle prior," *IEEE Trans. Image Process.*, vol. 32, pp. 2879–2888, 2023.

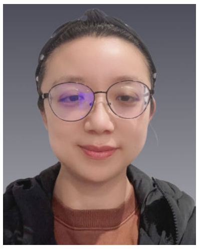
[23] F. Martínez, A. J. Rueda, and F. R. Feito, "A new algorithm for computing Boolean operations on polygons," *Comput. Geosci.*, vol. 35, no. 6, pp. 1177–1185, 2009.

[24] P.-E. Sarlin, D. DeTone, T. Malisiewicz, and A. Rabinovich, "SuperGlue: Learning feature matching with graph neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 4937–4946.

[25] S. Liu, L. Yuan, P. Tan, and J. Sun, "Bundled camera paths for video stabilization," *ACM Trans. Graph.*, vol. 32, no. 4, p. 78, Jul. 2013.

RUIFANG PAN was born in Nanchang, Jiangxia, in November 1959. She received the master's degree. She is currently the Dean of the School of Information, Zhejiang Guangsha Vocational and Technical University of Construction, and the Project Leader of the China Foreign Cultural Exchange Full Media Integration Base under the Ministry of Education. She is a second-level Professor. Her research interests include digital media integration research and virtual reality.

YUN ZHANG received the bachelor's and master's degrees from Hangzhou Dianzi University, in 2006 and 2009, respectively, and the Ph.D. degree from Zhejiang University, in 2013. He is currently a Professor with the Communication University of Zhejiang. From February 2018 to August 2018 and from December 2022 to October 2023, he was a Visiting Scholar with Cardiff University, U.K. His research interests include computer graphics, image and video editing, and computer virtual reality. He is a Senior Member of CCF.



LIN XU received the bachelor's, master's, and Ph.D. degrees from Fujian Normal University, in 2006, 2010, and 2014, respectively. She is currently pursuing the Ph.D. degree with the University of South Australia, in 2024. She is also a Lecturer with the Zhejiang Guangsha Vocational and Technical University of Construction, China. She visited the College of Creativity and Technology, Fo Guang Univeristy, in 2015, and Faculty of Sciences, Universite Libre De Bruxelles, in 2017. Her research interests include computer vision, multi-media systems, and computer intelligence.

AIHONG QIN received the Ph.D. degree in computer science and technology from Zhejiang University, in 2007. She is currently an Associate Professor with the Communication University of Zhejiang, China. Her research interests include computer graphics, video editing, and computer vision. She is a member of CCF.

HUI DU received the Ph.D. degree from the State Key Laboratory of CAD & CG, Zhejiang University, in 2013. He is currently a Faculty Member of the Communication University of Zhejiang. His research interests include image processing and computer vision.