• Review •

A survey on image and video stitching

Wei LYU, Zhong ZHOU*, Lang CHEN, Yi ZHOU

*State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering,
Beihang University, Beijing 100191, China*

* Corresponding author: zz@buaa.edu.cn

Received: 15 October 2018 Accepted: 20 November 2018

Supported by the National Natural Science Foundation of China (61872023).

Citation: Wei LYU, Zhong ZHOU, Lang CHEN, Yi ZHOU. A survey on image and video stitching. *Virtual Reality & Intelligent Hardware*, 2019, 1(1): 55—83
DOI: 10.3724/SP.J.2096-5796.2018.0008

Abstract Image/video stitching is a technology for solving the field of view (FOV) limitation of images/videos. It stitches multiple overlapping images/videos to generate a wide-FOV image/video, and has been used in various fields such as sports broadcasting, video surveillance, street view, and entertainment. This survey reviews image/video stitching algorithms, with a particular focus on those developed in recent years. Image stitching first calculates the corresponding relationships between multiple overlapping images, deforms and aligns the matched images, and then blends the aligned images to generate a wide-FOV image. A seamless method is always adopted to eliminate such potential flaws as ghosting and blurring caused by parallax or objects moving across the overlapping regions. Video stitching is the further extension of image stitching. It usually stitches selected frames of original videos to generate a stitching template by performing image stitching algorithms, and the subsequent frames can then be stitched according to the template. Video stitching is more complicated with moving objects or violent camera movement, because these factors introduce jitter, shakiness, ghosting, and blurring. Foreground detection technique is usually combined into stitching to eliminate ghosting and blurring, while video stabilization algorithms are adopted to solve the jitter and shakiness. This paper further discusses panoramic stitching as a special-extension of image/video stitching. Panoramic stitching is currently the most widely used application in stitching. This survey reviews the latest image/video stitching methods, and introduces the fundamental principles/advantages/weaknesses of image/video stitching algorithms. Image/video stitching faces long-term challenges such as wide baseline, large parallax, and low-texture problem in the overlapping region. New technologies may present new opportunities to address these issues, such as deep learning-based semantic correspondence, and 3D image stitching. Finally, this survey discusses the challenges of image/video stitching and proposes potential solutions.

Keywords Image stitching; Video stitching; Panoramic stitching; Registration; Alignment; Mesh optimization; Deep learning; 3D stitching

1 Introduction

Image stitching is among the oldest and most widely used topics in computer vision and graphics. In recent years, stitching algorithms have been applied in many fields (e.g., image processing, computer vision, and multimedia) and closely associated with the daily lives of people, such as constructing beautiful panoramas with smartphones applications, creating wide field-of-view (FOV) videos for surveillance, and assisting automobiles. Many well-known applications, such as Adobe Photoshop, AutoStitch¹, PTGui², and Image Composite Editor (ICE)³ effectively stitch multiple overlapping images to generate a wide-angle view. Meanwhile, various 360-degree polydioptric cameras based on panoramic stitching have been released, e.g., Nokia OZO⁴, GoPro Odyssey⁵, Facebook Surround 360⁶, and Samsung Gear360⁷. They construct a panorama from a sequence of images, and the panorama can be displayed with Virtual Reality (VR) head-mounted devices. However, restricted acquisition environments and datasets are required by these applications and cameras, and the various datasets to be practically stitched are nonstandard and contain many potential flaws, e.g., wide baseline, large parallax, changes in illumination and contrast, low texture, and occlusion.

Image stitching algorithms construct a wide-FOV view from a sequence of images by performing a basic pipeline consisting of three stages^[1]. In the first stage, the corresponding relationships between the original images are established by pre-calibrating the intrinsic and extrinsic parameters of the cameras^[2,3], or estimating a motion model based on pixels by calculating the optical flow^[4], per-pixel correspondence^[5-7], or sparse feature matching^[8-10]. In the next step, after the estimated transformations and registration is performed among the images, a projection plane is determined by selecting an image plane, e.g., the first image plane or an estimated intermediate plane, and then the registered images are deformed and aligned to the projection plane. Finally, the aligned images are fused together onto a large canvas by blending the same corresponding pixels in the overlapping regions between images and preserving the pixels in the non-overlapping regions. Most image stitching algorithms hypothesize that the original images are captured by a camera rotating about its optical center (e.g., most panoramic stitching algorithms^[11]), or that the scene is approximately planar (i.e., no parallax or minimal parallax), and violating these hypotheses results in inaccurate image registration, further misalignment, and ghosting. Multi-homography estimation and alignment optimization algorithms^[10,12,13] have been introduced to solve these issues, whereas robust and effective stitching algorithms are yet to be proposed.

Compared to image stitching, video stitching attracts less attention, probably owing to its direct relationship with image stitching. Video stitching is in many ways an extension or generalization of multi-image stitching, and large degrees of independent motion, camera zoom, and the desire to visualize dynamic events impose additional challenges. Most studies show that the original videos are captured in one of two modes, namely, static and dynamic (i.e., fixed optical sensor and moving cameras), and different issues and solutions correspond to each mode. Video stitching algorithms generally involve three steps^[14]: (1) A stitching template is first constructed by stitching the selected frames of original videos with image stitching algorithms; (2) A single wide-angle video is generated by stitching subsequence frames according

¹ <http://matthewalunbrown.com/autostitch/autostitch.html>

² <http://www.ptgui.com/>

³ <https://www.microsoft.com/en-us/research/project/image-composite-editor/>

⁴ <https://ozo.nokia.com/>

⁵ <https://gopro.com/odyssey>

⁶ <https://facebook360.fb.com/facebook-surround-360>

⁷ <https://www.samsung.com/global/galaxy/gear-360/>

to the template; (3) Foreground detection is employed to solve the potential blurring and ghosting in the stitched video, i. e., the stitching template is updated when an object moves across the overlapping regions among images (Figure 1). In contrast to fixed optical sensors, some studies focus on stitching videos captured with moving cameras^[15-17] (e. g., hand-held mobile cameras and unmanned aerial vehicles (UAVs)), and such videos usually pose additional challenges for stitching, e. g., shakiness and large parallax. Stabilization algorithms^[18,19] are usually adopted to eliminate the jittering artifacts in the stitched video.

Panoramic stitching has been developed along with the image/video stitching, and can be considered as a re-extension of image/video stitching, whereby a sequence of images/videos is stitched together in a closed-loop manner by using multi-image stitching algorithms, and projected to a cylinder or sphere, to generate a 360-degree panoramic view/video^[20,21]. In contrast to classical image / video stitching, it requires more restricted acquisition environments, i. e., the camera generally rotates 360 degrees about its optical center, which provides a good condition for adopting simpler stitching methods. Currently, numerous practical applications on stitching are based on panoramic stitching, such as combining structured camera arrays with simple stitching methods^[21], generating panoramas for smartphone applications, or panoramic cameras for surveillance. Meanwhile, owing to the requirements of these applications for high efficiency and great effectiveness, and the innovation of the hardware devices, more precise devices are combined with simpler stitching algorithms, accompanied by more pre-processing, e. g., a pre-calibration method is adopted to calibrate the parameters of the vision sensors^[22] in order to construct a constant stitching template to generate panoramas.

Because the basic theories of image stitching have been illuminated in research^[1], such as basic motion models of 2D motion, 3D transformations, cylindrical and spherical coordinates, and lens distortions, as well as the overview of early literatures, this survey mainly focuses on describing, classifying, and reviewing the influential studies on stitching in recent years, and analyzing and exploring potential issues and corresponding solutions. Meanwhile, some studies introduce the semantic understanding of scenes and 3D reconstruction techniques into the stitching field, such as deep learning-based semantic matching and 3D stitching.

We believe that it is time to revisit the research on stitching. The existing stitching methods perform very well on standard datasets and carefully chosen images, whereas practical datasets cannot be stitched as expected. Therefore a bottleneck has emerged and must be broken. The aim of this survey is to increase the attention paid to image/video stitching, and to foster in-depth thinking on the potential issues in the

Figure 1 Pipeline of stitching: (a) Input images/videos. (b) Calculating the similarity measure between pixels or feature points to register selected image pairs. (c) Deforming and aligning the registered images. (d) Blending the aligned images. Object detection should be performed to update the stitching template during (panoramic) video stitching.

field. More solutions are expected to be proposed by reviewing the research on stitching, which will further promote the research progress in this field.

This survey is organized as follows. We first provide an overview of the stitching algorithms in Section 2. The image stitching algorithms are grouped and classified according to the algorithm category and technological development in Section 3. Section 4 reviews video stitching approaches. Based on image and video stitching, panoramic stitching techniques are introduced in Section 5. According to the reviews of image, video, and panoramic stitching algorithms, we discuss the challenges facing the stitching field and their potential solutions in the final section.

2 Overview of image/video stitching algorithms

Image stitching algorithms focus on registering, aligning, and blending multiple overlapping images to generate a wide-FOV view. These are mainly divided into two categories: pixel-based methods, and feature-based methods.

Pixel-based methods (also known as direct methods), including most early works on stitching, register images by directly minimizing pixel-to-pixel dissimilarities^[7]. Many meaningful works have been developed for stitching, and utilizing the image information based on pixels, such as gradient, color, depth, and geometry, they estimate a global transformation model to deform and align images (e.g., a 3D rotation matrix). For example, because the gradient information is sensitive to high-level features in the image (e.g., lines, contours, and edges), some image stitching algorithms employ a coarse-to-fine strategy used until now by solving a cost function to optimize the approximate alignment in the gradient domain^[5-7]. Furthermore, the depth and color of pixels are utilized^[23], and some other works adopt the optical flow method and formulate the stitching as a manifold projection, whereby images are divided into many strips according to the calculated optical flow vector^[4]. Moreover, a graph structure has also been adopted to assist the stitching^[24]. These methods effectively register the images by taking advantage of image information, but require complex preprocessing and calculations, and they are limited to addressing images with simple scenes contained in the same plane.

Feature-based methods, a different class of algorithms proposed and widely used until now, adopt the sparse feature descriptors and feature matching implemented sequentially^[25]. Different from direct methods, feature-based methods are faster by selectively extracting sparse feature points rather than all pixels in the overlapping regions, more robust by effectively constructing different feature descriptors, and more highly automated by automatically calculating the adjacency relationships between input images. For more robust, faster, and more efficient performance, sparse feature matching is widely used in the stitching field (e.g., SIFT^[25], Harris^[26], SURF^[27], FAST^[28]), and multiple transformation matrixes (e.g., homography and affine) are simultaneously calculated in the stitching process. Similar to direct methods, early feature-based methods also deform and align images globally according to the estimated transformation^[8,9]. Furthermore, an image plane is divided into two dominant planes, including foreground and background planes^[10], or more planes^[12], and each plane corresponds to a transformation. These methods warp the images without distinguishing between overlapping and non-overlapping regions, which limits the algorithms to the simple scenes.

To address more complicated scenes, global transformations are insufficient, and local transformation methods have been proposed for such cases^[13,29]. An image is first divided into uniform grids, and each grid is warped and aligned with a homography estimated by introducing a Moving Direct Linear Transformation (Moving DLT) method in as-projective-as-possible (APAP)^[13]. Motivated by APAP, several

researchers have tried to stitch challenging datasets with parallax and baseline (i.e., natural baseline and wide baseline) by introducing mesh-based alignment optimization algorithms to improve the initial alignment. For example, Chang et al. warp the images according to different estimated transformations, where the warping changes from projective transformation to similarity transformation between the overlapping regions and non-overlapping regions^[30]. Lin et al. focus on optimizing the seam in the stitching process^[31]. Zhang and Liu^[31] first roughly align the input images by performing SIFT^[25] and RANSAC^[33], and then construct an objective function consisting of alignment and some prior constraint terms to further optimize the alignment. It effectively stitches the images with wide-baseline, and manual interaction is required⁸.

Based on image stitching, video stitching combines object detection and stabilization algorithms to construct a wide-FOV video. For the static cameras, a classical video stitching model initially stitches the selected frames of input videos to generate a stitching template, after which the subsequence frames are stitched according to the template, and the template is updated when objects moves across the overlapping regions between videos^[14]. Other works focus on stitching videos captured by moving cameras (e.g., hand-held mobile cameras, or UAVs), which combine video stitching and stabilization to construct a wide-FOV video and eliminate the jitter artifacts^[15,16,34]. Moreover, Lin et al.^[17] introduce a simultaneous localization and mapping method^[35] to calculate the camera motions, and a 3D reconstruction to reconstruct the overlapping regions.

Panoramic stitching is closed-loop stitching, i.e., each image is registered with the corresponding images in a sequence of images, and the registered images are deformed, aligned, and projected onto a cylindrical or spherical surface. Most works on panoramic stitching estimate the camera parameters with general perspective transformation methods, perform some optimization methods to reduce the cumulative error, and then blend the aligned images to seamlessly generate a 360-degree panoramic view/video^[20,36]. Certain studies also construct unstructured or structured camera arrays to simulate a 360-degree camera rotation^[37]. Similar panoramic devices associated with VR have also been released in recent years, such as Facebook Surround 360 (Figure 2). Pre-calibrating the intrinsic camera parameters and bundle adjustment (BA)^[9] are adopted to adjust the extrinsic parameters of cameras to rectify the camera positions, after which the stereoscopic parallax is calculated and a panorama is constructed by using an optical flow method. Figure 3 shows the classification of image and video stitching.

Numerous reports show that most recent studies on stitching adopt a coarse-to-fine strategy to stitch images/videos, and the performance of both SIFT and RANSAC has been demonstrated. Although many other methods have been proposed (e.g. SURF, Harris, FAST), these two methods are the most widely used for registration and alignment, and then mesh-based alignment optimization methods are introduced to further improve the alignment. They generally perform very well on standard datasets or carefully chosen images with consistent exposure and a camera rotating about its optical center, whereby a sequence of images is stitched together to generate a high-quality view without ghosting, blurring, or other faults.

Figure 2 Facebook Surround 360. There are 14 Wide-Angle Lenses horizontally distributed on the side, one Fish-eye Lens vertically distributed on the upper part, and two Fisheye Lenses vertically distributed under the device body.

⁸ <http://www.cad.zju.edu.cn/home/gfzhang/projects/panorama/pano-video.wmv>

Figure 3 Classification of image and video stitching.

However, many images to be stitched are very different from the standard datasets and bring additional challenges to stitching, such as wide baseline, parallax, and low texture. To determine wide baseline and parallax more clearly, different camera spacings (i.e., baselines) of the images are defined^[38], a median distance between adjacent images of 0.8 m is regarded as a natural baseline, 1.6 m is regarded as a wide baseline, and 2.4 m is regarded as a very wide baseline. The degree of image parallax is closely related to the camera baseline, and the presence of the baseline leads to parallax, e.g., normal human vision uses parallax to estimate the distances to objects in daily life. The two "eyes" represent two viewpoints. By looking at a nearby object and alternatively blinking each of the eyes, an offset between the two positions observed by the two viewpoints is found, and the distance between two viewpoints is called the baseline. Parallax fades away as the distance to the object increases, and it disappears when the distance is infinite or the baseline is zero. The effect of the baseline gradually decreases as the distance increases. The baseline in the standard datasets is generally the natural baseline or smaller, no parallax or minimal parallax is required, large parallax leads to blurring and ghosting, and wide baseline leads to mismatching and misalignment (Figure 4).

Currently, the performance of most stitching methods is still insufficient for many practical applications, such as automobiles and video surveillance. These practical applications generally pre-calibrate the intrinsic and extrinsic camera parameters^[3,36], and adopt simpler methods, i.e., replacing the complex algorithms with higher cost. To explore potential solutions, some works introduce deep learning-based semantic matching^[39–43]. Dense features are learned and extracted from images using Convolutional Neural Networks (CNNs), and the learned features are more flexible than handcrafted features (e.g., SIFT, Harris) and take advantage of the image information. Finally, this paper introduces a 3D stitching method combining 3D reconstruction and image stitching^[44].

3 Image stitching algorithms

According to the algorithm category and technological development, this section groups and classifies image stitching algorithms into two categories: pixel-based (direct) methods, and feature-based methods. The development of image stitching is also divided into two stages according to the new requirements introduced by practical applications and the innovation of various technologies.

In the first stage, a large number of works were presented to basically satisfy the needs of wide-FOV

Figure 4 Three main challenges brought by practical applications: (a) Wide baseline. (b) Large parallax. (c) Low texture. The left column shows the original images, and the stitched results are shown in the right column.

photos for better exhibition in photography. The original photos were strictly captured by professionals using professional cameras. Direct methods were widely used initially, and the efficiency and effectiveness were sufficient for professional applications. Furthermore, many sparse feature descriptors were introduced to match images more efficiently. In particular, the performance of the SIFT^[25] descriptor has been demonstrated in many researches. These feature-based methods general combine global alignment, BA^[9], and a multi-band blending algorithm to construct wide-FOV views^[8,9], and also produce some effective applications, e.g., AutoStitch and ICE. The second breakthrough of image stitching was obtained in the second stage, along with faster technological innovation and various requirements. After a mesh-based multi-homography method was proposed^[13], numerous algorithms have focused on optimizing alignment, and many effective optimization schemes have been proposed. In this section, we review pixel-based methods and feature-based methods that have been used for image stitching. Figure 5 shows the classification of image stitching.

3.1 Pixel-based (direct) methods

To take advantage of image information (e.g., intensity, gradient, color, and geometry), pixel-based (direct) methods register multiple images by directly minimizing pixel-to-pixel dissimilarities. This section classifies the direct methods into different methods based on manifold projection, gradient domain, depth information, and graph structure.

3.1.1 Manifold projection-based methods

Global cumulative error occurs in the process of stitching a sequence of partial images, when the camera tilts while rotating or moving. The size of the aligned images gradually decreases, and even tends to a point as the number of original images increases.

To solve this issue, Peleg et al. formulated image stitching as aligning strips extracted from the original

Figure 5 Classification of image stitching.

images^[4]. The direction of the optical flow is orthogonal to the axis of the selected strips, and all images are projected onto an envelope of image planes, i. e., stitching is performed by projecting thin strips from the images onto manifolds (Figure 6). Because the directions of the optical flow are parallel to each other in the images, and the errors are assigned to each image, the global cumulative error is effectively avoided. Meanwhile, more restricted motion models are employed to produce final views in a very fast stitching process, albeit with less accuracy.

Figure 6 Alignment between registered strips from three images. (a), (b), and (c) Strips are orthogonal to the optical flow. (d) The registered strips are deformed and aligned.

3.1.2 Gradient domain-based methods

Because gradient information is sensitive to high-level features in images (e.g., lines, contours, and edges), and conducive to the understanding of image scenes, Levin et al. constructed several different stitching methods based on a gradient domain image stitching (GIST) approach^[5], where each method corresponded to a cost function, and the quality of different functions was evaluated and compared^[6]. The preferred method is the L1 optimization of a feathered cost function on the original image gradients (GIST1), owing to overcoming the geometric misalignments. They minimize the dissimilarity measure between the derivatives of the stitched image and the derivatives of the input images, and the images are registered, aligned, and blended in the gradient domain rather than the intensity domain, and seam artifacts and edge duplications are effectively reduced.

Jia and Tang^[7] only registered one or two optimal partitions to construct a sparse set of deformation vectors detected as a 1D sparse feature in the partition areas, and smoothly propagated the deformation into the target image by minimizing the cost in the gradient domain to align the image structures and intensities.

3.1.3 Graph-based methods

Uyttendaele et al. constructed a graph to stitch images when some objects move across the overlapping regions^[24]. The vertices of the graph represent the regions of difference (RODs) in the overlapping regions and the edges link the corresponding RODs (Figure 7). Higher weights are assigned to larger and more central ROD vertices, to selectively remove all but one instance of each object, avoiding the motion discontinuities of objects arising from either of the selected side images. Each image is divided into patches, with each patch corresponding to a calculated quadratic transfer function. The functions in each

patch are averaged with those of their neighbors, and further adjustments are made for local variations in exposure, then the transferred pixels are blended with the corresponding transfer function.

Bujnák and Sara also constructed an oriented graph. All possible correspondences are defined as vertices, edges represent the potential registration relationship between images (i.e., matching uniqueness)^[45], and the orientation of edges represents better pairwise correspondence based on matching quality^[46].

3.1.4 Depth-based methods

In contrast with the^[24] construction of a graph to stitch images containing moving objects, a depth-based method (DBM) combines a camera projection model and depth estimation method, and the virtual stitching viewpoint is coincident with either of the inputs^[23]. The color and depth of each pixel are first calculated by performing a plane sweep algorithm^[47] and graph cut optimization^[48], to stitch the overlapping regions. To synthesize the non-overlapping regions, the color is segmented, the depth is propagated to adjacent color segments, and the smooth appearance connection among them is preserved in the stitched views. Furthermore, the static image stitching method is extended to stitch videos. The original videos are divided into different layers using a mixture of Gaussians approaches, such as foreground and background layers, and then these layers are projected onto a stitching plane according to the respective depth estimation.

3.1.5 Conclusion and comparison

Direct methods mainly focus on addressing the issues caused by the properties of the image itself, such as difference in brightness, while pixel-based matching is generally inefficient. Existing direct methods are all limited to address images with a single plane, or the scene is roughly planar, and there is no parallax. More details are shown in Table 1.

Table 1 Comparison of different direct methods

Approach	Principal	Moving object	Pre-processing	Advantage	Weakness
Peleg et al. ^[4]	Optical flow			Fast. Address cumulative error.	Low accuracy.
Levin et al. ^[5]	Gradient weighting		√	Seamless.	Input aligned images.
Jia and Tang ^[7]	Locally aligning			Precise alignment.	Distortion.
Uyttendaele et al. ^[24]	Graph structure	√		Eliminate ghosting caused by moving objects.	Complicated calculation.
Zhi and Cooperstock ^[23]	Depth and color	√		Address a certain degree of depth discontinuity.	Complicated calculation.

3.2 Feature-based methods

In contrast to direct methods, feature-based methods estimate a 2D motion model with sparse feature points. First, we briefly review the local feature descriptors. Lowe presented a sparse feature descriptor known as Scale-invariant feature transform (SIFT), which remains widely used to date^[25]. SIFT is mainly used to detect and describe image features. It is invariant to image translation, rotation, and scale, and robust to change in 3D viewpoint, addition of noise, affine distortion, and changes in illumination and contrast. Numerous studies have shown that the SIFT is the most widely used feature descriptor in image

Figure 7 Regions of difference (RODs). (Upper) Alignment among three images with a moving object. (Bottom) Corresponding RODs caused by moving object in the original images.

stitching and the performance has been demonstrated. SIFT mainly consists of four stages: scale-space extrema detection, keypoint localization, orientation assignment, and keypoint descriptor (Figure 8).

Figure 8 Construction of SIFT descriptor^[25]. (a) Image pyramid. (b) Extrema detection for DOG pyramid. (c) Creation of keypoint descriptor.

In the first stage, an image pyramid is constructed by repeatedly convolving input images with Gaussians, including a set of scale-space images shown on the left in Figure 8(a), and the adjacent Gaussian images are subtracted to produce a difference-of-Gaussian (DOG) pyramid shown on the right in Figure 8(a). In the second stage, the maxima and minima locations of the DOG images are detected by comparing a pixel to its neighbors in 3×3 regions at the current and adjacent scales shown in Figure 8(b). The next stage assigns one or more orientations to each detected keypoint, based on the local gradient directions shown on the left in Figure 8(c). Finally, the keypoints are defined by calculating the gradient magnitude and orientation at each sample point around the keypoint areas shown on the right in Figure 8(c).

There are also some other feature descriptors. For example, the Harris corner detector calculates the intensity difference between adjacent pixels to detect the salient regions, i.e., corners, edges, and smooth areas, although it is not invariant to image scale^[26]. Based on SIFT, Speeded-up robust features (SURF)^[27] introduce a Hessian matrix-based measure for the detector to obtain faster feature extraction, which is mainly suitable for camera self-calibration. Similar to Harris, Features from accelerated segment test (FAST) selects a keypoint when there are continuous pixels around a circle, and these pixels are all brighter than the intensity of the candidate pixel minus a threshold^[28].

Our survey of image/video stitching shows that almost all influential feature-based stitching algorithms adopt the SIFT descriptor to extract feature points from the original images. According to the issues and development in the stitching field rather than simple methodological classification, we mainly group and classify the feature-based stitching methods into global single transformation, and local hybrid transformation methods (Figure 9).

3.2.1 Global single transformation methods

Global transformation methods deform and align images by adopting the same transformation models, e. g., projective or affine, and these methods do not distinguish the overlapping and non-overlapping regions among the original images. Some methods estimate a single transformation model^[8,9] for alignment. Furthermore, other methods divide an image plane into multiple planes, with each plane corresponding to a transformation^[10,12]. They perform well on images with simple scene structures, and the overlapping and non-overlapping regions undergo similar transferring, which provides the basis for mesh-based alignment methods.

A milestone approach was proposed by Brown and Lowe^[8,9], adopting object recognition techniques based on local invariant features to match images. Image stitching is formulated as matching multiple images with a sparse feature descriptor, and a probabilistic model is used for verification^[8], and gain compensation and automatic straightening are used to further improve the stitched view^[9], known as the well-known AutoStitch. It estimates a global homography transformation to effectively deform and align images, while strictly assuming that the overlapping regions must be laid on the same plane, i. e., the baseline between cameras is a natural baseline or smaller baseline. AutoStitch introduces the sparse feature matching into image stitching, and the sparse feature-based methods have dominated image stitching.

To satisfy more complicated applications and solve the issue of aligning the images containing multiple planes in the overlap regions, Gao et al. presented a dual-homography transformation model, and an image was divided into two dominant planes, such as the foreground plane and background plane, with each plane corresponding to an estimated homography transformation^[10]. Furthermore, an image was divided into multiple planes^[12], with each plane corresponding to an affine transformation. A local optimal affine stitching field is first calculated, and then a smooth field with better extrapolation ability is created. It is flexible enough to address parallax while retaining the good extrapolation, to effectively address occlusion, and further to avoid local optimal and obtain better alignment.

3.2.2 Local hybrid transformation methods

The mesh-based alignment method was a breakthrough for image stitching initially. The images are first divided into uniform meshes (Figure 10), with each mesh corresponding to an estimated transformation, and an increasing number of studies have presented different optimization strategies, mainly including hybrid transformation models and various prior constraints. The overlapping and non-overlapping

regions undergo different processes according to different transformations. For example, the non-overlapping regions are generally warped by using a similarity transformation model to avoid potential distortions, the overlapping regions are warped and aligned with an estimated projection transformation, and then a smooth field was introduced to smooth the area between two regions.

Zaragoza et al. (APAP) divided images into uniform meshes, with each mesh corresponding to a homography transformation estimated with a Moving DLT method^[13]. They first introduces the mesh-based

Figure 9 Classification of feature-based stitching.

Figure 10 Mesh-based image deformation and alignment.

alignment scheme into the image stitching, locally preserves the salient structures in the non-overlapping regions, and reduces the distortions caused by warping in the overlapping region. Owing to the excellent performance on aligning the overlapping images, APAP is considered as one of the state-of-the-art methods until now, and numerous studies works on stitching are based on APAP. For example, Liu and Chin^[29] proposed a data-driven warp adaptation scheme to improve APAP^[13], and further identified the misaligned regions and inserted appropriate points, i.e., increasing the flexibility of the warp, and improving the alignment. Furthermore, the research on image stitching has entered the next stage, i.e. local hybrid transformation based on mesh alignment optimization. This section groups and classifies local hybrid transformation methods into hybrid extrapolation-based, seam-based optimization, mesh-based alignment optimization, and high level feature-based methods.

(1) Hybrid extrapolation-based methods

To obtain a more natural stitched view, Chang et al. (SPHP) adopted a local hybrid transformation model^[30]. The projective transformation of the overlapping regions was smoothly extrapolated to the non-overlapping regions, and the warping gradually changed from projective to similarity across the two regions. The salient structures are preserved with a similar constraint and the image distortions are simultaneously reduced, although this is not suitable for overlapping regions with multiple planes. Similarly, a smooth stitching field over the target image is introduced to linearize the homography transformation^[49], and it is changed to the global similarity transformation to mitigate the perspective distortion in the non-overlapping regions. However, the estimation of the global similarity transformation is not robust to unnatural rotation and scaling, and locally produces distortions during stitching.

Li et al. employed a quasi-homography warp scheme to balance the perspective distortion against the projective distortion in the non-overlapping regions^[50]. The images are first divided into two spaces with a vertical line: the overlapping and non-overlapping regions separately belong to different spaces. Then, warping is formulated as solving a bivariate system, perspective distortion and projective distortion are respectively characterized as slope preservation and scale linearization, the original images are warped according to a general homography, and an estimated quasi-homography is used to second-warp the images.

(2) Seam-based optimization methods

Zhang and Liu constructed a hybrid transformation model to handle images with a certain degree of parallax^[51]. An optimal homography is first estimated to roughly align the images, a content-preserving warp method^[52] is adopted to further improve the alignment, and then a seam-driven method^[53] and multi-band blending are employed to obtain a final panorama. Effective and robust stitching is performed to address the images with a certain degree of parallax, while the warping hypotheses are searched in a randomized fashion, and the current process does not take advantage of the alignment knowledge generated from the previous iterations.

Based on Zhang and Liu's work^[51], Lin et al.^[31] improved the seam-driven method^[53] by adjusting feature weights according to their distances to a candidate seam. A superpixel-based feature clustering method is first performed to generate different alignment hypotheses, and then optimal seams are iteratively generated according to different feature weights. Meanwhile, a structure-preserving warp method is adopted to preserve curves and straight structures (Figure 11). It obtains significant performance gains with more calculation cost.

To take advantage of the ability of the Deep Learning framework, Herrmann et al.^[32] leveraged the

Figure 11 Pipeline of the seam-guided stitching algorithms^[31].

advances in object detection^[54], and combined the stitching algorithm^[55] to construct an object-centered image stitching framework. The method focuses on searching for an optimal seam with some new constraint terms. Different from the multi-registration method^[55], a single registration is adopted in the seam finding stage^[32], and a cropping term penalizing seams that cuts through an object with the cost proportional to the length of the seam. Moreover, a duplication term avoids duplication, and an occlusion term promoting the occlusion label. However, the method is limited to stitch images containing detectable objects.

(3) Mesh-based alignment optimization methods

To reduce the distortions, a coarse-to-fine stitching scheme is introduced^[56]. Initially, registration and alignment between images are obtained by performing APAP^[13], and then the alignment optimization is formulated as solving an objective function consisting of different constraints. These include an alignment term ensuring the alignment quality after warping, a local similarity term serving for regularization and propagating alignment constraints from the overlapping regions to the non-overlapping regions, and a global similarity term ensuring each warped image undergoes a similarity transformation. To be more robust to rotation and scaling, the BA method is adopted to estimate the focal length and 3D rotation of each image^[9].

Zhang et al. presented a pure mesh-based optimization alignment framework with a coarse-to-fine manner, to optimize alignment and regularity in 2D space^[57]. A local homography verification method is adopted to roughly align images, and various prior constraints are adopted to improve the alignment with an iterative optimization scheme. It uses a shape correction method to stitch images with wide baseline. Meanwhile, many works adopt similar methods of designing more prior constraints, while a certain coupling relationship between different constraints results in local optimization rather than global optimization, and high computational cost. Similar to this process^[57], Herrmann et al. calculated a homography to initially register parts of images, refined its mesh representation, and then calculated an optimal seam by solving a cost function^[55].

The following example can be used to illuminate such methods: a grid, whose four vertices are denoted as v_1, v_2, v_3, v_4 respectively, contains a feature point P_n , with corresponding to the deformed grid, i.e., its vertices are v'_1, v'_2, v'_3, v'_4 , and the feature point P'_n , as illustrated in Figure 12.

Figure 12 Feature point interpolation. (Left) Original grid and a feature point P_n . (Right) The deformed vertices and feature point P'_n .

The Interpolation weights of the four vertices are:

$$\begin{aligned}
 \omega_1 &= (v_{3,x} - P_{n,x})(v_{3,y} - P_{n,y}) \\
 \omega_2 &= (P_{n,x} - v_{4,x})(v_{4,y} - P_{n,y}) \\
 \omega_3 &= (P_{n,x} - v_{1,x})(P_{n,y} - v_{1,y}) \\
 \omega_4 &= (v_{2,x} - p_x)(P_{n,y} - v_{2,y})
 \end{aligned} \tag{1}$$

Where $(P_{n,x}, P_{n,y})$ is the two-dimensional coordinate of the feature points P_n , $(v_{1,x}, v_{1,y})$ is the two-dimensional coordinate of the four vertices.

The alignment term is:

$$E_A(V) = \sum_{(p'_i, q'_i) \in P} \|p'_i - q'_i\|^2 = \sum_{(p'_i, q'_i) \in P} \|W_i V - W_j V\|^2 \tag{2}$$

where P is a set of correspondences from all image pairs, p'_i, q'_i are the deformed positions of the matched points respectively, whose coordinates are the weighted sums of the mesh vertices in V . W_i, W_j are a set of the interpolation weights of the four grid vertices neighboring p, q respectively.

The scale preserving term is:

$$E_S(V) = \sum \|S(I'_i) - \mu \cdot S(I_i)\|^2 \tag{3}$$

where the original images, I_i , corresponding to the i -th deformation, I'_i , and S is a scale measurement for images defined as a 2D vector, μ is a content.

The objective function is:

$$E(V) = \lambda_A E_A(V) + \lambda_S E_S(V) \tag{4}$$

where V is a 2m dimension vector storing the coordinates of vertices, and λ_A, λ_S are the weight factors of every term. Furthermore, some other constraints are widely adopted, such as line constraint for preserving the linear structure, and smoothness constraint for enabling that the grid takes a similarity transformation to preserve the salient structures of overlapping regions.

(4) High-level feature-based methods

Owing to different preferences of photographers, low texture regions objectively appear in some images, and increase the difficulty of feature detection. Xiang et al. took advantage of the high-level features of images, i.e., line and point features were used to obtain more matches from images with low-texture overlapping regions^[58]. The extracted line features are used to guide the local warping, and are further optimized by solving an objective function, i.e., containing point alignment term, global alignment term, smoothness term, and line constraint term consisting of line correspondence term and line collinearity term. A global similarity constraint is adopted to adjust the local warping, and reduce the projective distortion in the non-overlap regions. Besides, it constructs mesh vertices in the overlap regions as the set of candidate points to match regions with insufficient matched features. The line and contour information are used to improve the alignment^[31]. These methods mitigate the issue of low texture, and the key solutions are to take advantage of the image information to obtain enough matches. For example, an approximate nearest-neighbor algorithm (approximate NNF) is presented to search for approximate nearest-neighbor matches between image patches, known as PatchMatch^[59]. This is an image matching method, which could be further used for image stitching.

3.2.3 Conclusion and comparison

Feature-based methods have dominated image stitching, and work well on standard datasets and some challenging datasets. These works are mainly concerned with obtaining better results than higher efficiency, with the result in that they are still not suitable for practical applications. Table 2 shows a

Table 2 Comparison of different feature-based methods

Approach	Principal	Parallax	Advantage	weakness
AutoStitch ^[9]	Simple Sparse feature matching.		Automatic.	Limited to single plane.
Gao et al. ^[10]	Dual-homograph.		Two planes.	Single homography.
Lin et al. ^[12]	Multi-affine.	✓	Address small parallax.	Single affine.
Zaragoza et al. ^[13]	Mesh-based.		Multiple transformations.	Local distortion.
Liu and Chin ^[29]	Inert more appropriate correspondence based on pixels.	✓	Address the images with a certain degree of parallax, and wide-baseline.	Complicated calculation.
Chang et al. ^[30]	Local hybrid transformation model.		Different transformations between overlapping and non-overlapping regions.	Limited to parallel or vertical images.
Li et al. ^[50]	Quasi homography.		Reduce distortion.	Limited to parallel images.
Zhang and Liu ^[51]	Hypotheses of warping.	✓	Effective and robust stitching.	The alignment knowledges are not reusable.
Lin et al. ^[31]	Seam-driven.	✓	Optimization seam, reduce ghosting.	Iterative operations and complicated calculation.
Herrmann et al. ^[32]	Prior constraints.		Single registration, and address the images containing moving objects.	Limited to images containing detectable objects.
Chen and Chuang ^[56]	Coarse-to-fine scheme.	✓	Scale and rotation correction.	Local distortion.
Zhang et al. ^[37]	Various prior constraints and optimization terms.		Address the wide-baseline images.	Complicated calculation, local distortion.
Xiang et al. ^[58]	High-level features.		Address the images with a certain degree of low texture.	Local distortion.

detailed comparison of these methods. Early feature-based methods^[9,10,12,13] deform and align the images with a transformation method (homograph or affine). They do not distinguish between overlapping and non-overlapping regions, leading distortion and ghosting, and warp the images by applying different transformations to overlapping and non-overlapping, and perform well on limited images^[30,49,50]. Furthermore, other methods focus on optimizing the stitching seam, and also obtain better results, although along with complicated calculations^[31,32,51]. Finally, mesh-based alignment optimization methods are introduced to address challenging datasets^[56,57], and some methods belonging to other categories also adopt these methods^[32,58]. These methods design various prior constraints to optimize the alignment, although the local distortion still cannot be avoided owing to the coupling relationships between different constraints.

4 Video stitching algorithms

Video stitching attracts less attention than image stitching, probably owing to their relationship: they are both similar and different. Video stitching is in many ways an extension and generalization of multi-image stitching, and large degrees of independent motion, camera zoom, and the desire to visualize dynamic events impose additional challenges, i.e., jitter and shakiness. This section classifies video stitching methods into those based on static cameras and moving cameras according to the video acquisition manner.

4.1 Static cameras

A classical video stitching system focuses on stitching images captured by static cameras and an object detection method is performed. The selected frames of the input videos are first stitched to generate a stitching template, in which there is no object moving across the overlapping regions between the input videos. The subsequence frames are in turn stitched according to the template, and then the template is updated when there are moving objects.

Rav-Acha et al. embedded the detected dynamic contents into the stitched images with a stitching algorithm^[4] and an object detection step^[60]. Much consistent information is contained in the frames of

spatially adjacent videos captured by a panning camera, and a precise alignment is obtained.

He and Yu performed a coarse-to-fine stitching process for surveillance applications^[14]. The selected frames of two input videos are first divided into different layers, and the backgrounds are stitched with a conventional stitching pipeline when there are no moving objects across the overlapping regions^[1]. Similarly, the matched feature pairs are clustered to multiple layers involving different objects, where each layer contains a set of matched feature pairs according to the same homography, and the videos are overall pre-aligned^[10]. To avoid missing information, ghosting, and artifacts caused by moving objects, the variations in gradient are calculated in the overlap region, and to further update the optimal seam.

Different from the classical process, Jiang and Gu^[61] formulated a spatio-temporal mesh optimization framework based on Zhang and Liu's method^[51]. All frames of the input videos are initially aligned according to the estimated spatial and temporal global transformations, and an objective function is constructed to represent the matching costs in the spatial domain and temporal domain, to be solved to improve the geometric alignment. Similar to video texture synthesis^[62], higher weights are assigned to salient regions such as spatial edges and temporal edges, to preserve the salient structures in the videos and search for an optimal seam.

4.2 Moving cameras

Some works focus on stitching videos captured by moving cameras (e.g., a smartphones or UAVs), which brings stitching additional challenges (i.e., jitters and shakiness). The methods such as^[18,19] can remove jittery and shaky motions, but cannot be directly applied to stabilize the stitched videos. To solve the jitter in the stitched video, many researches generally perform image stitching and stabilization simultaneously.

Guo et al.^[34] simultaneously performed video stitching and stabilization. Two transformations were estimated, i.e., the inter-transformation between different cameras to obtain the spatial alignment, and the intra-transformation within each video to maintain the temporal smoothness. Meanwhile, a mesh-based warping method is adopted for alignment^[13], and a bundled-paths approach is used as a baseline for video stabilization, to synthesize a smooth virtual camera path^[19]. It effectively stitches the scenes with a certain degree of parallax, assuming that the camera stably moves while being limited to a certain degree of freedom, initially enforcing a rough synchronization, and misalignments are caused by large depth variations and strong motion blur^[34].

Su et al. formulated video stitching as solving an objective function consisting of a stabilization term and a stitching term, and performed an iterative optimization scheme^[15]. Recently, Nie et al. extended the method^[15]. The backgrounds of input videos are first identified with a background identification method, and to be stitched, and then a false match elimination scheme is introduced to reduce mismatches^[16]. Finally, a scoring strategy is introduced to evaluate the stabilization quality.

Lin and Liu combined a dense 3D reconstruction and camera pose estimation technique to stitch videos captured by hand-held cameras^[17]. The CoSLAM system was first adopted to recover 3D camera motions and sparse scene points^[35], reconstructing 3D scenes in the overlapping regions, constructing a smooth virtual camera path staying in the middle of all the original paths^[52], and then a mesh-based warping optimization method called Line-Preserving Video Warp (LPVW) method is adopted to synthesize the stitched video along the path.

4.3 Conclusion and comparison

Video stitching on static cameras is generally performed according to the process shown in Figure 1, and stabilization is required for stitching on moving cameras (Table 3). Meanwhile, exiting video stitching works mentioned above all fail to balance the results and efficiency, i.e., better results accompanied by complicated calculations.

Table 3 Comparison of feature-based methods

Approach	Principal	Stabilization	Advantage	Weakness
Rav-Acha et al. ^[60]	Embed the detected objects into the stitched backgrounds		Accurate alignment	Limited to the videos captured by a panning camera.
He and Yu ^[14]	Layer stitching		Accurate alignment	Complicated calculation.
Jiang and Gu ^[61]	Spatio-temporal mesh optimization		Improve the geometric alignment	Complicated calculation.
Guo et al. ^[34]	Spatio-temporal alignment	✓	Address the scenes with a certain degree of parallax	Limited to the cameras with a certain degree of freedom and require pre-processing.
Su et al. ^[15]	Optimization function	✓	Balance the stitching and stabilization	Complicated calculation.
Nie et al. ^[16]	Discriminatively stitching the videos between backgrounds and foregrounds	✓	Improve the matching in ^[46]	Complicated calculation.
Lin et al. ^[17]	Estimate parameters of the cameras	✓	Accurately calculate the 3D camera paths	Limited to the videos with-depth continuity

5 Panoramic algorithms

In recent years, with the development of VR and stitching techniques, various panoramic stitching-based 360-degree polydioptric cameras have been released. Different from a single-fisheye camera directly obtaining 180-degree wide-angle views, they construct panoramas based on panoramic stitching methods, and stitch multiple overlapping images/videos captured by a 360-degree distributed camera array. Meanwhile, panoramic stitching is a re-extension of the image/video stitching and works well for a camera rotating 360-degree about its optical center. Many studies combine unstructured or structured camera arrays and stitching algorithms to create panoramas, to be directly displayed with VR devices. Some representative methods will be listed.

5.1 A single camera

A 360-degree panorama is constructed when original images are captured by a camera rotating about its optical center. However, it practically depends on the amount of camera translation relative to the nearest objects in front of the camera^[20].

Szeliski and Shum estimated only three unknowns in the rotation matrix instead of the eight parameters in general perspective transformations, to deform and align a sequence of images is captured by a hand-held camera rotating 360-degree about its optical center^[20]. The images at both the beginning and ending of the sequence are matched together, to simultaneously update and adjust the estimated focal lengths and rotation matrices after constructing a panorama, and to further handle cumulative error^[63]. Then, the 3D rotation matrix (and optionally, focal length) associated with each image is intrinsically more stable than estimating a full eight-degree-of-freedom homography, which makes this a method of choice for large-scale image stitching algorithms^[1]. It divides each image into small patches and estimates patch-based alignments and provides guidance to the later algorithms^[13,29,30]. Duffin and Barrett adopted a spherical arc length instead of angles to parameterize the positions of a panoramic image^[11]. The angles were used to position each candidate image on the common spherical center, and the strong coupling between the focal length and the angles was mitigated.

To avoid matching all pairs of frames, Steedly et al. assumed that most temporally adjacent frames were spatially adjacent, and only searched for the matches between all pairs of temporal adjacent images^[64]. Similarly, key frames are selected by locally matching the overlapping regions, and the degree of overlap is strongly limited to a threshold range (i.e., 25% to 50%), and the beginning and ending frames are defined as key frames^[6].

According to the preferences of the artists and common users, regular rectangular boundaries give better visual experience in the panoramas. He et al.^[65] extended the seam carving algorithm^[66] to construct rectangular images from panoramic images. A horizontal or vertical seam was first inserted through the input image and expanded the seam by one pixel horizontally or vertically, the irregular panoramic image was warped to a rectangle by displacing all pixels on one side of each seam, and then the rectangular image was meshed. Finally, a mesh-based global optimization step was employed to preserve the perceptual properties, such as shapes and straight lines.

Similar to the general RANSAC method using four-point for homography estimation^[33], Brown et al. proposed a modified three-point RANSAC algorithm to estimate rotational motion models, introducing a semi-calibrated strategy and all the parameters other than the focal length were known^[36]. Furthermore, Kaynig et al. extended the three-point solver^[36] to construct an image system, image processing was formulated as an ideal pinhole projection plus radial distortion, and lens distortions were effectively corrected^[2]. The proposed method^[2] is much faster, though more expensive than the method in Brown et al.'s work^[36]. To obtain faster stitching, Silva et al.^[3] performed stitching on GPUs with SURF^[27] and RANSAC^[33]. Pre-calibrated meshes can be loaded, and some parameters can be manually modified to accommodate fine details. For real-time processing, the algorithm is limited to use faster processing methodologies, thereby losing some certain effects and requiring more cost.

5.2 Camera arrays

Different from panoramic stitching based on a single camera, several studies construct camera arrays to simulate single camera motion that rotates about its optical center (Figure 2). Because these approaches are generally devoted to practical applications, more real-time and simpler algorithms are required.

Perazzi et al. extended the local warping method based on optical flow^[7], and constructed unstructured camera arrays without the accurate placement of cameras (Figure 13)^[37]. To establish an optimal sequence of pairwise deformations, an iterative maximum weighted graph matching^[67] is adopted to search for the corresponding set of input views, and each image is independently warped with a minimum parallax error. The method^[37] reduces hardware costs with a better algorithm, whereas the jitters cannot be solved, assuming the configuration of cameras to be static over the input sequence.

Figure 13 Unstructured camera arrays^[37].

Lee et al. presented a 360-degree panoramic stitching system for videos and constructed a structured camera array known as Rich360^[21]. First, pre-calibrating the tightly structured camera rigs, i.e., the intrinsic and extrinsic parameters of the cameras are robustly calibrated with a classical calibration^[22], and then the result is reused as a template for video stitching. Second, the input images are individually projected onto a projection surface with varying depth according to the corresponding parameters, and the disparities in the overlapping regions are minimized. Then, the stitched spherical panoramas are fitted into a rectangular frame. The importance of a region is measured by combining the image gradient, saliency, and face

detector, and more rays (i.e., pixels) are assigned to an important region, to fully exploit the resolution of the source videos. Compared to Perazzi and Sorkine-Hornung's method^[37], Rich360 strongly requires tightly structured camera arrays and less flexibility.

Pan et al. introduced a panoramic stitching scheme into the automobile system called Rear-Stitched View Panorama (RSVP)^[68]. RSVP assigns four cameras around a vehicle, two cameras on both sides to replace the side external rear-view mirrors, and a stereo pair of cameras on the rear to replace the center interior rear-view mirror. The extrinsic camera parameters are pre-calibrated (i.e., location and orientation) with a precisely designed set of three calibration charts with known dimensions, and then the videos, which are captured by both side cameras and the left stereo camera, are stitched together to generate a seamless panorama of the rear surroundings. Meanwhile, the right stereo camera is used to estimate depth, the intrinsic and extrinsic parameters are all used to project original images to a virtual camera view, and then an optimal seam is determined to generate a spatially-coherent single RSVP view.

5.3 Fisheye lens camera arrays

Owing to the wide FOV and much more information contained in a 180-degree fisheye view, some methods construct panoramas from a sequence of images captured by fisheye lens camera arrays. Agarwala et al. construct multi-viewpoint panoramas of long image scenes with a coarse-to-fine process, and a translating hand-held fisheye lens camera is used to capture the original images^[69]. However, it has insufficient automation and greater limitations compared to other methods, e.g., more pre-processing and manual processing are required. In the pre-processing stage, it removes the radial distortion using available software⁹, compensates for the exposure variation, and recovers the camera projection matrices using SIFT^[25] and a structure-from-motion method^[70]. Besides, an image surface is manually defined as a projection plane.

In 2017, Samsung Research America released a Gear 360 camera consisting of two fisheye lens cameras whose FOV is close to 195 degrees each^[71]. The corresponding stitching approach compensates for the intensity fall-off of fisheye lens cameras, unwraps the fisheye images, registers and aligns them together with an adaptive alignment method, and then blends the aligned images. Different from the methods^[21,37], it^[71] replaces unstructured/structured camera arrays with a dual-fisheye lens camera structure. Because a single fisheye lens has a wider FOV than an ordinary lens camera, a dual-fisheye lens stitching has a smaller cumulative error. However, the images captured by fisheye lens cameras have the limited overlapping FOV, and misalignment between two lenses gives rise to visible discontinuities in the stitching boundaries.

5.4 Conclusion and comparison

In contrast to image and video stitching, panoramic stitching methods focus on stitching a standard 360-degree sequence of images/videos in a closed-loop manner. Generally, panoramic stitching is a technique that could be used in practical applications. It adopts the pre-calibrated methods^[2,3,36] to calibrate the parameters of cameras, then combines the stitching algorithms to generate the final panoramas. These works all perform very well on panoramic stitching combined with some certain devices, such as structured camera arrays. Meanwhile, some works also perform very well on panoramic stitching, but they require particular acquisition environments or standard datasets^[11,20,64]. Finally, other work can address a certain degree of parallax caused by unstructured camera arrays, though this is limited to static camera configurations^[37].

⁹ EPAPERPRESS, 2005. (<http://epaperpress.com/ptlens/>)

6 Challenges and extensions

This paper has provided a detailed survey and overview of the excellent research on image/video stitching presented in recent years. The development of image stitching is mainly divided into two stages. Pixel-based methods and early feature-based methods globally deform and align multiple overlapping images according to an estimated single transformation. The methods presented in recent years (e.g., most feature-based methods based on mesh optimization alignment) perform a local hybrid transformation model to warp and align the images. Under the latter methods, salient structures are better preserved, and more natural stitched views are obtained. Panoramic stitching generally assumes that the reasonable acquisition of datasets is required to mitigate the limitations of algorithms (i.e., a camera rotates 360-degree about its optical center, or this is simulated with camera arrays), and pre-calibration methods are widely adopted to obtain faster stitching. Though stabilization should be used in some situations (i.e., moving cameras), video stitching is based on image stitching, more effective and robust image stitching is required for better video stitching.

6.1 Challenges

Image stitching has been researched for decades and numerous methods have been proposed. Most methods only perform well on standard datasets (i.e., natural baseline, and little or no parallax) (Figure 14), and some methods try to solve the more challenging datasets (e.g., wide baseline and parallax) by introducing more complex algorithms. Meanwhile, most practical applications in daily life prefer to simplest solutions, e.g., complex devices are combined with the simplest algorithms such as pre-calibrating the cameras. Numerous studies have shown that practical datasets to be stitched are more complex, and the current methods are insufficient (Figure 15).

Figure 14 Image stitching on standard datasets.

Figure 15 Image stitching on challenging datasets. Most methods based on 2D stitching fail to stitch four images with wide baseline in the right column, and the 3D stitching method effectively fuses the images at the bottom.

The practical images to be stitched impose certain challenges, as follows:

(1) Wide baseline and large parallax

In contrast with professional acquisition technology (e.g., photography), practical images tend to be captured in a more flexible and casual way, e.g., the images are casually captured with a portable device (i.e., a mobile phone or digital camera) in daily life. The angle and exposure difference between cameras result in

many issues, e.g., wide baseline, large parallax, and differences in brightness.

(2) Low-texture overlapping regions

Low-texture regions objectively appear in some images. Large proportions of background regions are contained in the scenes of images, such as a granite floor with the same pattern in an artificial scene, or a single landscape in a natural scene (e.g., Sky, Flower Sea, Lake, and Forest).

(3) Very wide baseline and very large parallax

Stitching has been widely used in modern surveillance systems. Very wide baseline is caused by the irregular distribution of surveillance cameras. The baselines among the cameras may be a few meters, or more than 10 m. For example, the cameras are distributed in parallel across a square (i.e., airport lobby), which also results in low overlap among images. Meanwhile, the distance from cameras to the targets is generally relatively short, resulting in very large parallax.

Essentially, Parallax introduces ghosting and blurring to stitching, whereas wide baseline and low-texture overlapping regions between two cameras cause the issue that too few matched feature points are detected, such that the transformations between images cannot be accurately calculated, which leads to misregistration.

6.2 Potential solutions

Numerous studies have shown that the challenges of wide baseline, large parallax, and low-texture overlapping regions can probably be solved in the 2D stitching field. These issues mainly result in the inability to obtain enough matched feature pairs, further misregistration for stitching, or ghosting and blurring in the stitched views. To solve these issues, some works provide references as follows:

(1) Mesh-based alignment optimization methods are still the potential solutions to stitch multiple images with wide baseline, whereas simply more different prior constraints are far from enough to solve the essential issues. A sufficient number of matched feature points and more accurate registration is required, but not per-pixel matching as in direct methods.

(2) High-level feature-based methods utilize high-level information of images, such as lines, edges, and contours, to obtain more matches from images with low texture, though it is limited to artificial scenes (e.g., standard divided ground structures with regular texture).

(3) Seam-based optimization methods could also be adopted to address images with large parallax, although many estimation operations of alignment and seam are iteratively performed, and the computational complexity is generally not accepted.

Some works understand the contents in the images by learning the features from images by using CNNs^[72]. The learned features are more flexible than traditional handcrafted features (e.g., SIFT), and take advantage of the image information. CNN features could be adopted to search for more salient features from images, and combine optimization schemes to improve the registration and alignment.

Furthermore, very wide baseline leads to low overlap among images. In such cases, the transformations among images cannot be effectively calculated, and inconsistent depth is caused by very large parallax. These do not belong to the 2D stitching field, and automatic matching cannot be performed with 2D stitching algorithms. To fuse the images with these issues, 3D reconstruction is introduced, and a 3D stitching algorithm combining 3D reconstruction and 2D stitching is proposed^[44].

To effectively solve the presented issues, this survey introduces three potential solutions as follows:

(1) Coarse-to-fine stitching scheme with deep learning-based semantic matching

The learned CNN features are more flexible, and more potential matched candidates could be extracted

from images with wide baseline or low-texture regions. The initial alignment is obtained with semantic matching, and then the alignment is improved by combining a reasonable optimization strategy (e.g., mesh-based alignment optimization methods).

(2) Layering-based seam optimization combined with semantic segmentation

Ghosting and blurring are generally caused by large parallax, i.e., there is a position offset when observing the same object through different viewpoints, and a shorter distance to the object corresponds to a larger offset. An image could be divided into different layers, such as the background and foreground, where each layer corresponds to a semantic label, and a disparity map could be estimated with stereo matching. The backgrounds could be stitched separately, the position of foreground objects could be blended together according to the disparity map, followed by blending the background and foreground.

(3) 3D stitching combining 3D reconstruction and 2D stitching

Automatic registration and alignment could not be performed for images with very wide baseline and very large parallax. The initial image registration and alignment on a 3D scene could be obtained with 3D reconstruction from a single image, and 2D stitching is used to improve the alignment.

Some images with more complex structures and transformations cannot be fused with 2D stitching or 3D stitching. In such cases, methods from other fields could be adopted, such as 3D reconstruction and simultaneous localization and mapping (SLAM), and point cloud-based methods are introduced to register the images. We hope that the several potential schemes presented above are significant to promote further research on image stitching and fusion. The detailed principle of semantic matching and 3D stitching methods is as follows:

(1) Deep learning-based semantic matching for stitching

The superior performance of deep learning has been proved in recent years, and many works understand the semantic contents of images with learned CNN features. Owing to their invariant to geometric deformations and changes in illumination, CNN features accurately locate the salient features. Deep learning-based semantic image matching may be a potential solution for the challenges (i.e., wide baseline and low-texture) faced by stitching.

Figure 16 shows the basic pipeline of semantic correspondence, in which CNN features learned by using pre-trained CNNs are adopted rather than handcrafted features. Two mainly schemes using for matching: 1) End-to-end methods take advantage of the superior of deep learning architectures, but some prior information such as labeled data are required. 2) Preprocessing-based methods initially learn CNN features from the images, then the correspondences are obtained with some post-processing strategies.

Ufer and Ommer^[39] adopted the pre-trained Alexnet on ImageNet dataset^[40] to generate image pyramids and feature pyramids, and extracted the CNN features from the feature maps of the first four convolution layers according to the Shannon formula and non-maximum suppression algorithm. To obtain the optimal correspondences, they solved an objective function consisting of similarity constraints, and geometric constraints utilizing the geometrical information (i.e., position, direction, and distance). Similarly, Aberman et al. adopted a pre-trained VGG-

Figure 16 Pipeline of semantic correspondence. Given image pair (I_1, I_2) , feature extraction is performed by using pre-trained CNNs models, matching the learned CNN features, and then warping and aligning the images according to an estimated transformation.

19 network to generate a five-layer feature map pyramid, constructed a feature descriptor with nominal dimension for each neuron, and then searched for the Neural Best Buddies pairs with a nearest neighbor matching algorithm^[41]. A coarse-to-fine scheme is introduced to precisely locate the matched feature pairs from the high to low level.

Ignacio Rocco et al.^[42] constructed an end-to-end CNN framework to simulate the matching process with SIFT^[23], to be trained on labeled datasets in a highly supervised manner. The loss function is the mean square error of labeled feature points, and then the geometric transformation with several freedom parameters is outputted. The affine transformation and thin-plate spline transformation respectively corresponds to two modules in the framework, which improves the alignment accuracy and fault tolerance.

The pyramidal affine regression network (PARN) estimates locally-varying affine transformation fields across semantical similar images in a coarse-to-fine manner^[43]. Firstly, it estimates a global affine transformation over an entire image, then progressively increases the degree of freedom of the transformation in the form of a quad-tree, and finally produces pixel-wise continuous affine transformation fields.

These studies provide potential schemes to solve the issues of wide baseline and low-texture in image stitching, i.e., a coarse-to-fine strategy combining the semantic understanding of images and 2D stitching methods, semantic-based matching could be adopted to roughly align the images with similar semantic objects, and then some optimization methods could be introduced to improve the alignment.

(2) 3D stitching

To effectively fuse the images with very wide baseline and very large parallax, Zhou et al. proposed a 3D stitching method based on 3D reconstruction from a single image, by estimating reference camera poses and 2D image matching information to align their adjacent 3D models^[44]. Matched point pairs or selected keypoint pairs are added into the intersecting plane space of 3D models, and dense mesh vertices are constructed with the added point pairs, to further deform the original images. The aligned images are then blended by selecting a virtual viewpoint. 3D models play an important role in the process of image matching, transformation, and fusion, and the pose parameters of cameras and reconstructed models are required to project and back-project the selected point pairs. The pipeline of 3D stitching is shown in Figure 17.

Figure 17 Pipeline of 3D stitching.

This method performs well on a suitable depth plane, and each image is segmented into multiple sub-regions according to different depths caused by very large parallax. Matching and warping are performed to reduce mismatches and large distortions. The final stitched results can be observed at any viewpoint without a large distortion, though there are some disadvantages as follows: First, the effectiveness of 3D stitching is limited by interactive reconstruction, and an automatic depth estimation method is yet to be introduced. Second, the final results are three-dimensional and too flexible, and distortion is effectively reduced, while limiting user to observe all the information by rotating the angle of view.

Currently, various requirements and the innovation of different technologies bring new opportunities and challenges to image/video stitching, and the stitching technique is already indispensable in daily life and professional applications. For example, large amounts of data containing videos and images are collected by city-level surveillance systems, and stitching could be as an aid to organize and manage numerous redundant data, i.e., constructing a panoramic urban surveillance system, or for urban traffic management. Meanwhile, the stitching technique can also be used in artificial intelligence (AI) systems, e.g., generating

360-degree panoramic video around a vehicle in automobiles, and cross-camera object tracking in stitched panoramas. However, the presented stitching methods are far from sufficient, more effective and robust algorithms are required. We believe that many fields will be changed by the breakthroughs of stitching.

7 Conclusion

We have conducted a comprehensive survey of existing approaches on image and video stitching and introduced different approaches according to the algorithm category and technological development. Some challenges introduced by practical applications are discussed, and some thoughts for future research are given. As an increasingly useful topic in various fields, the research on stitching should focus more on practical applications than on experimental datasets. We hope our survey will increase the attention given to stitching and motivate new research efforts to advance the field of research on image and video stitching.

References

- 1 Szeliski R. Image Alignment and Stitching: A Tutorial. Foundations and Trends® in Computer Graphics and Vision, 2007; 2(1): 1–104
DOI:10.1561/0600000009
- 2 Kaynig V, Fischer B, Buhmann J M. Probabilistic image registration and anomaly detection by nonlinear warping. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Anchorage, Alaska, USA, 2008, 1–8
DOI:10.1109/CVPR.2008.4587743
- 3 Silva R M A, Gomes P B, Frensh T, Monteiro D. Real time 360° video stitching and streaming. In: ACM SIGGRAPH 2016 Posters. Anaheim, California: ACM, 2016: 1–2
DOI:10.1145/2945078.2945148
- 4 Peleg S, Rousso B, Rav-Acha A, Zomet A. Mosaicing on adaptive manifolds. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000, 22(10): 1144–1154
DOI:10.1109/34.879794
- 5 Levin A, Zomet A, Peleg S, Weiss Y. Seamless image stitching in the gradient domain. In: Computer Vision-ECCV 2004: 2004// 2004. Berlin, Heidelberg. Springer Berlin Heidelberg, 2004, 377–389
- 6 Zomet A, Levin A, Peleg S, Weiss Y. Seamless image stitching by minimizing false edges. IEEE Transactions on Image Processing, 2006, 15(4): 969–977
DOI:10.1109/TIP.2005.863958
- 7 Jia J, Tang C K. Eliminating structure and intensity misalignment in image stitching. In: Tenth IEEE International Conference on Computer Vision (ICCV'05). Beijing, China, 2005, 1651–1658
DOI:10.1109/ICCV.2005.87
- 8 Brown M, Lowe D G. Recognizing Panoramas. In: Proceedings of the IEEE International Conference on Computer Vision. Nice, France, 2003, 3: 1218
- 9 Brown M, Lowe D G. Automatic panoramic image stitching using invariant features. International Journal of Computer Vision, 2007, 74(1): 59–73
DOI:10.1007/s11263-006-0002-3
- 10 Gao J, Kim S J, Brown M S. Constructing image panoramas using dual-homography warping. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Colorado Springs, CO, USA, 2011, 49–56
DOI:10.1109/CVPR.2011.5995433
- 11 Duffin K L, Barrett W A. Fast focal length solution in partial panoramic image stitching. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Kauai, HI, USA, 2001, II–II
DOI:10.1109/CVPR.2001.991031
- 12 Lin W Y, Liu S, Matsushita Y, Ng T T, Cheong L F. Smoothly varying affine stitching. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Colorado Springs, CO, USA, 2011, 345–352
DOI:10.1109/CVPR.2011.5995314

- 13 Zaragoza J, Chin T J, Brown M S, Suter D. As-projective-as-possible image stitching with moving DLT. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Portland, OR, USA, 2013, 2339–2346
DOI:10.1109/CVPR.2013.303
- 14 He B, Yu S. Parallax-Robust Surveillance Video Stitching. 2016, 16(1):7
DOI:10.3390/s16010007
- 15 Su T, Nie Y, Zhang Z, Sun H, Li G. Video stitching for hand-held inputs via combined video stabilization. In: SIGGRAPH ASIA 2016 Technical Briefs. Macao, China, 2016: 25
- 16 Nie Y, Su T, Zhang Z, Sun H, Li G. Dynamic Video Stitching via Shakiness Removing. IEEE Transactions on Image Processing, 2018, 27(1): 164–178
DOI:10.1109/TIP.2017.2736603
- 17 Lin K, Liu S, Cheong L F, Zeng B. Seamless video stitching from hand-held camera inputs. Computer Graphics Forum, 2016, 35(2): 479–487
DOI:10.1111/cgf.12848
- 18 Matsushita Y, Ofek E, Weina G, Xiaoou T, Heung-Yeung S. Full-frame video stabilization with motion inpainting. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2006, 28(7): 1150–1163
DOI:10.1109/TPAMI.2006.141
- 19 Liu S, Yuan L, Tan P, Sun J. Bundled camera paths for video stabilization. ACM Trans. Graph. 2013, 32(4): 1–10
DOI:10.1145/2461912.2461995
- 20 Szeliski R, Shum H Y. Creating full view panoramic image mosaics and texture-mapped models. In: SIGGRAPH. Los Angeles, 1997, 251–258
- 21 Lee J, Kim B, Kim K, Kim Y, Noh J. Rich360: optimized spherical representation from structured panoramic camera arrays. ACM Transactions on Graphics (TOG), 2016, 35(4): 1–11
DOI:10.1145/2897824.2925983
- 22 Zhang Z. A flexible new technique for camera calibration. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2000, 22: 1330–1334
- 23 Zhi Q, Cooperstock J R. Toward dynamic image mosaic generation with robustness to parallax. IEEE Transactions on Image Processing, 2012, 21(1): 366–378
DOI: 10.1109/TIP.2011.2162743
- 24 Uyttendaele M, Eden A, Skeliski R. Eliminating ghosting and exposure artifacts in image mosaics. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Kauai, HI, USA, 2001, II–II
DOI:10.1109/CVPR.2001.991005
- 25 Lowe D G. Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision, 2004, 60(2): 91–110
DOI:10.1023/B:VISI.0000029664.99615.94
- 26 Harris C, Stephens M. A combined corner and edge detector. In: Proceedings of Alvey vision conference. Manchester, UK, 1988
- 27 Bay H, Ess A, Tuytelaars T, Van Gool L. Speeded-Up Robust Features (SURF). Computer Vision and Image Understanding, 2008, 110(3): 346–359
DOI:10.1016/j.cviu.2007.09.014
- 28 Rosten E, Drummond T. Machine Learning for High-Speed Corner Detection. In: Computer Vision – ECCV 2006. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, 430–443
- 29 Liu W X, Chin T J. Correspondence Insertion for As-Projective-As-Possible Image Stitching. 2016
- 30 Chang C H, Sato Y, Chuang Y Y. Shape-preserving half-projective warps for image stitching. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Columbus, OH, USA, 2014, 3254–3261
- 31 Lin K, Jiang N, Cheong L-F, Do M, Lu J. Seagull: seam-guided local alignment for parallax-tolerant image stitching. In: Computer Vision – ECCV 2016. Cham: Springer International Publishing, 2016, 370–385
- 32 Herrmann C, Wang C, Bowen R S, Keyder E, Zabih R. Object-Centered Image Stitching. In: Computer Vision – ECCV 2018. Springer International Publishing, 2018, 846–861
DOI:10.1007/978-3-030-01219-9_50

- 33 Fischler M A, Bolles R C. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 1981, 24(6): 381–395
DOI:10.1145/358669.358692
- 34 Guo H, Liu S, He T, Zhu S, Zeng B, Gabbouj M. Joint Video Stitching and Stabilization From Moving Cameras. *IEEE Transactions on Image Processing*, 2016, 25(11): 5491–5503
DOI:10.1109/TIP.2016.2607419
- 35 Zou D, Tan P. Coslam: Collaborative visual slam in dynamic environments. *IEEE transactions on Pattern Analysis and Machine Intelligence*, 2013, 35(2): 354–366
DOI:10.1109/TPAMI.2012.104
- 36 Brown M, Hartley R I, Nistér D. Minimal solutions for panoramic stitching. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Minneapolis, Minnesota, USA, 2007, 1–8
DOI:10.1109/CVPR.2007.383082
- 37 Perazzi F, Sorkine-Hornung A, Zimmer H, Kaufmann P, Wang O, Watson S, Gross M. Panoramic video from unstructured camera arrays. *Computer Graphics Forum*. 2015, 34(2): 57–68
DOI:10.1111/cgf.12541
- 38 Flynn J, Neulander I, Philbin J, Snaveley N. Deepstereo: Learning to predict new views from the world's imagery. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Las Vegas, Nevada, 2016, 5515–5524
- 39 Ufer N, Ommer B. Deep semantic feature matching. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Honolulu, HI, USA, 2017, 5929–5938
DOI:10.1109/CVPR.2017.628
- 40 Deng J, Dong W, Socher R, Li L, Kai L, Li F-F. ImageNet: A large-scale hierarchical image database. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Miami, Florida, USA 2009, 248–255
DOI:10.1109/CVPR.2009.5206848
- 41 Aberman K, Liao J, Shi M, Lischinski D, Chen B, Cohen-Or D. Neural Best-Buddies: Sparse Cross-Domain Correspondence. *ACM Transactions on Graphics (TOG)*, 2018, 37(4): 69
DOI:10.1145/3197517.3201332
- 42 Rocco I, Arandjelovic R, Sivic J. Convolutional neural network architecture for geometric matching. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Honolulu, HI, USA, 2017, 39–48
DOI:10.1109/CVPR.2017.12
- 43 Jeon S, Kim S, Min D, Sohn K. PARN: Pyramidal Affine Regression Networks for Dense Semantic Correspondence Estimation. In: *Proceedings of European Conference on Computer Vision*. Munich, Germany, 2018, 355–371
DOI:10.1007/978-3-030-01231-1_22
- 44 Zhou Y, Cao M, You J, Meng M, Wang Y, Zhou Z. MR Video Fusion: Interactive 3D Modeling and Stitching on Wide-baseline Videos. In: *ACM Symposium on Virtual Reality Software and Technology*. Tokyo, Japan: ACM, 2018, 1–11
DOI: 10.1145/3281505.3281513
- 45 Bujnák M, Sara R. A robust graph-based method for the general correspondence problem demonstrated on image stitching. In: *2007 IEEE 11th International Conference on Computer Vision*. Rio de Janeiro, Brazil, 2007, 1–8
DOI: 10.1109/ICCV.2007.4408884
- 46 Šára R. The principle of stability applied to matching problems in computer vision. RR CTU–CMP–2007–17. Center for Machine Perception, Czech Technical University, 2007
- 47 Collins R T. A space-sweep approach to true multi-image matching. In: *Proceedings CVPR IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. San Francisco, CA, USA, 1996, 358–363
DOI:10.1109/CVPR.1996.517097
- 48 Boykov Y, Kolmogorov V. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2004, 26(9): 1124–1137
DOI:10.1109/TPAMI.2004.60
- 49 Lin C C, Pankanti S U, Natesan Ramamurthy K, Aravkin A Y. Adaptive as-natural-as-possible image stitching. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Boston, MA, USA, 2015, 1155–1163
- 50 Li N, Xu Y, Wang C. Quasi-homography warps in image stitching. *IEEE Transactions on Multimedia*, 2018, 20(6): 1365

-1375

DOI:10.1109/TMM.2017.2771566

- 51 Zhang F, Liu F. Parallax-tolerant image stitching. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Columbus, OH, USA, 2014, 3262–3269
- 52 Liu F, Gleicher M, Jin H, Agarwala A. Content-preserving warps for 3D video stabilization. In: Acm Siggraph. 2009, 1–9
DOI:10.1145/1576246.1531350
- 53 Gao J, Li Y, Chin T J, Brown M S. Seam-Driven Image Stitching. In: Proceedings of Euro-Graphics. Girona, Spain, 2013, 45–48
- 54 Ren S, He K, Girshick R, Sun J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In: International Conference on Neural Information Processing Systems, 2015, 91–99
- 55 Herrmann C, Wang C, Bowen R S, Keyder E, Krainin M, Liu C, Zabih R. Robust Image Stitching with Multiple Registrations. In: Computer Vision – ECCV 2018. Munich, Germany: Cham: Springer International Publishing, 2018, 53–69
DOI:10.1007/978-3-030-01216-8_4
- 56 Chen Y S, Chuang Y Y. Natural Image Stitching with the Global Similarity Prior. In: Computer Vision - ECCV 2016. Cham: Springer International Publishing, 2016, 186–201
DOI:10.1007/978-3-319-46454-1_12
- 57 Zhang G, He Y, Chen W, Jia J, Bao H. Multi-Viewpoint Panorama Construction With Wide-Baseline Images. IEEE Transactions on Image Processing, 2016, 25(7): 3099–3111
DOI:10.1109/TIP.2016.2535225
- 58 Xiang T Z, Xia G S, Bai X, Zhang L. Image stitching by line-guided local warping with global similarity constraint. Pattern Recognition, 2018, 83:481–497
DOI:10.1016/j.patcog.2018.06.013
- 59 Barnes C, Shechtman E, Finkelstein A, Goldman D B. PatchMatch: a randomized correspondence algorithm for structural image editing. ACM Trans. Graph. 2009, 28(3): 1–11
DOI:10.1145/1531326.1531330
- 60 Rav-Acha A, Pritch Y, Lischinski D, Peleg S. Dynamosaics: video mosaics with non-chronological time. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Diego, CA, USA, 2005, 58–65
DOI:10.1109/CVPR.2005.137
- 61 Jiang W, Gu J. Video stitching with spatial-temporal content-preserving warping. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2015, 42–48
DOI:10.1109/CVPRW.2015.7301374
- 62 Kwatra V, Schödl A, Essa I A, Turk G, Bobick A F. Graphcut textures: image and video synthesis using graph cuts. ACM Transactions on Graphics (ToG), 2003, 22(3): 277–286
DOI:10.1145/1201775.882264
- 63 Shum H Y, Szeliski R. Construction and refinement of panoramic mosaics with global and local alignment. In: Proceedings of the IEEE International Conference on Computer Vision. Bombay, India, 1998, 953–956
DOI:10.1109/ICCV.1998.710831
- 64 Steedly D, Pal C, Szeliski R. Efficiently registering video into panoramic mosaics. In: Proceedings of the IEEE International Conference on Computer Vision. Beijing, China, 2005, 2: 1300–1307
DOI:10.1109/ICCV.2005.86
- 65 He K, Chang H, Sun J. Rectangling panoramic images via warping. ACM Transactions on Graphics (TOG), 2013, 32(4): 1–10
DOI:10.1145/2461912.2462004
- 66 Avidan S, Shamir A. Seam carving for content-aware image resizing. ACM Transactions on Graphics (TOG), 2007, 26(3): 10
DOI:10.1145/1276377.1276390
- 67 Galil Z. Efficient algorithms for finding maximal matching in graphs. In: Colloquium on Trees in Algebra and Programming. Berlin, Heidelberg: Springer Berlin Heidelberg, 1983, 90–113
- 68 Pan J, Appia V, Villarreal J, Weaver L, Kwon D. Rear-Stitched View Panorama: A Low-Power Embedded Implementation for Smart Rear-View Mirrors on Vehicles. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. Honolulu, HI, USA, 2017, 1184–1193

DOI:10.1109/CVPRW.2017.157

- 69 Agarwala A, Agrawala M, Cohen M, Salesin D, Szeliski R. Photographing long scenes with multi-viewpoint panoramas. *ACM Transactions on Graphics (TOG)*, 2006, 25(3): 853–861
DOI:10.1145/1141911.1141966
- 70 Brown M, Lowe D G. Unsupervised 3D object recognition and reconstruction in unordered datasets. In: *Proceedings of the International Conference on 3-D Digital Imaging and Modeling*. Ottawa, Ontario, Canada, 2005, 56–63
DOI:10.1109/3DIM.2005.81
- 71 Ho T, Budagavi M. Dual-fisheye lens stitching for 360-degree imaging. In: *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*. New Orleans, USA, 2017, 2172–2176
DOI:10.1109/ICASSP.2017.7952541
- 72 Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks. In: *Proceedings of Conference on Neural Information Processing Systems*. Lake Tahoe, Nevada, USA, 2012, 1097–1105