IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received April 4, 2018, accepted May 6, 2018, date of publication May 11, 2018, date of current version June 5, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2835659

Stitching for Multi-View Videos With Large Parallax Based on Adaptive Pixel Warping

KYU-YUL LEE AND JAE-YOUNG SIM, (Member, IEEE)
School of Electrical and Computer Engineering, Ulsan National Institute of Science and Technology, Ulsan 44919, South Korea

Corresponding author: Jae-Young Sim (jysim@unist.ac.kr)

This work was supported in part by the National Research Foundation of Korea within the Ministry of Science and ICT (MSIT) under Grant 2017R1A2B4011970 and within the Ministry of Education under Grant 2016R1D1A1A09919618, and in part by the Institute for Information and Communications Technology Promotion through the Korea Government (MSIT), Information-Coordination Technique Enabling Augmented Reality with Mobile Objects under Grant 20170006670021001.

ABSTRACT Conventional stitching techniques for images and videos are based on smooth warping models, and therefore, they often fail to work on multi-view images and videos with large parallax captured by cameras with wide baselines. In this paper, we propose a novel video stitching algorithm for such challenging multi-view videos. We estimate the parameters of ground plane homography, fundamental matrix, and vertical vanishing points reliably, using both of the appearance and activity-based feature matches validated by geometric constraints. We alleviate the parallax artifacts in stitching by adaptively warping the off-plane pixels into geometrically accurate matching positions through their ground plane pixels based on the epipolar geometry. We also exploit the inter-view and inter-frame correspondence matching information together to estimate the ground plane pixels reliably, which are then refined by energy minimization. Experimental results show that the proposed algorithm provides geometrically accurate stitching results of multi-view videos with large parallax and outperforms the state-of-the-art stitching methods qualitatively and quantitatively.

INDEX TERMS Multi-view videos, video stitching, image stitching, large parallax, adaptive pixel warping, epipolar geometry.

I. INTRODUCTION

Multi-view videos are widely used in many applications such as surveillance [1]–[3], sports [4]–[6], virtual training [7] and video conferencing [8], [9]. One of the essential techniques for multi-view applications is stitching, which combines multiple images, captured from different viewing positions and directions, to generate a single image with a wider field of view [10]. Image stitching has been actively studied in the literatures [11]–[21], and related commercial products have been also developed, e.g., Adobe Photoshop Photomerge^{TM} and Microsoft Image Composite Editor. Moreover, many current mobile devices with cameras are able to synthesize a panorama image by stitching multiple images captured at different time instances. Also, around view monitoring is one of the core applications of autonomous vehicles, which employs bird's eye views of stitched multiple images captured by front, side, and rear view cameras [22].

Traditional image stitching methods assume that a pair of images is taken from very close camera locations to each other and the captured scene structures are roughly planar. Based on these assumptions, we obtain stitched images by performing the three major steps: feature matching, image alignment, and image composition. First, feature points are detected from different images, which are then matched together by using feature descriptors, e.g., SIFT [23]. In the alignment step, a global image warping model such as homography is estimated by using the obtained feature matches, and multiple images are aligned to a common image domain accordingly. Finally, the pixel values in a stitched image are determined by average blending or seam cutting methods [10].

However, when multi-view cameras capture non-planar scene structures at relatively far camera positions from one another, resulting multi-view images exhibit parallax phenomenon where the relative locations of scene contents are varying across different views. In such cases, the traditional stitching methods suffer from parallax artifact. Therefore, advanced image stitching methods [11]–[21] have been studied which alleviate some amount of parallax artifact by designing locally adaptive transformations for flexible

warping, employing similarity transformation to reduce perspective distortion, and/or hiding the misalignment in composition stage based on seem-cutting method.

Recently, in many practical applications such as surveillance and sports, static multiple cameras are placed at very far viewing positions from one another with wide baselines. Also, captured 3D real-world scenes often include multiple foreground objects moving over a wide range of scene depths. For example, walking pedestrians are captured by static multiple cameras installed at arbitrary locations [24]–[26], and multiple players in sports games are captured by static cameras with wide baselines [27]. On these challenging multi-view images, even the aforementioned advanced image stitching techniques have limitations to combine the diverse scene structures accurately causing ghosting artifacts in stitching results due to the two main reasons. First, abrupt depth discontinuity among multiple foreground objects and background is hard to be treated accurately by the existing warping schemes. Second, appearance-based feature descriptors may provide large numbers of outlier matches due to severe parallax.

Compared to the image stitching research, relatively little effort has been made to develop multi-view video stitching techniques. Video stitching was regarded as an extension of image stitching where the multiple frames from different views at a certain time instance are stitched together by using existing image stitching techniques [28]. Also, a temporal cost term is simply added to the cost function for image stitching [29]. Therefore, stitching for challenging multi-view videos with large parallax still has the aforementioned problems of image stitching.

In this paper, we propose a geometrically accurate stitching algorithm for multi-view videos with large parallax (MVLP) which are captured by stationary cameras with wide baselines. We also consider surveillance and sports applications where multiple people are moving on the ground plane at arbitrary distances from the cameras. We develop a parallax-adaptive pixel warping model, where the ground plane pixels are warped by homography, but the pixels off the plane, i.e. the pixels on the foreground objects and the distant background region, are warped through their ground plane pixels based on the epipolar geometry. We also estimate the optimal ground plane pixels by employing both of the reliable spatial and temporal feature matches based on energy minimization framework. Experimental results show that the proposed algorithm stitches multi-view videos successfully without severe parallax artifacts, and yields a significantly better performance than that of the existing state-of-the-art image stitching techniques qualitatively and quantitatively.

A preliminary result of this work was presented in [30]. The major differences between [30] and this paper are as follows.

- We propose a more generalized video stitching framework which aligns the foreground objects and the background, respectively, while our previous algorithm [30] was applied to the foreground objects only.
- We improve the warping performance by estimating optimal ground plane pixels, while our previous work [30] estimates a projective depth using the lowest pixel in each object.
- We perform more extensive experiments using 12 video sequences and provide comparative experimental results between the conventional methods and the proposed algorithm qualitatively and quantitatively.

The rest of this paper is organized as follows. Section II describes the related work on image and video stitching and static multi-camera based tracking. Section III proposes the basic concept of the proposed parallax-adaptive pixel warping model. Section IV and Section V explain the algorithms of parameter estimation and ground plane pixel estimation, respectively. Section VI presents the experimental results. Finally, Section VII concludes the paper.

II. RELATED WORK

A. IMAGE AND VIDEO STITCHING

Homography is a traditional image warping model which describes the projective relationship between two image planes based on the planar scene assumption [10], [31]. In general, an optimal homography is estimated by feature matching between two images. Homography can register multiple images associated with small camera baselines successfully, however, it fails to work on the images with large camera baselines where a captured scene is composed of multiple planar structures.

To overcome this limitation, advanced image stitching methods employ spatially-varying warps which adaptively align spatial deviation between two images caused by parallax. Gao *et al.* [11] estimated dual homographies to align the ground plane and the distant background plane, respectively, by clustering the feature points according to their positions. Lin *et al.* [12] initialized a global affine transformation which is then iteratively refined to minimize a cost function defined by matched features. Zaragoza *et al.* [13] partitioned an input image into multiple cells, and estimated a homography for each cell by weighting feature matches according to the relative distances to the feature points. Zhang *et al.* [14] proposed a mesh-based alignment technique to mitigate the shrinking problem of wide-baseline panorama synthesis, which designs a scale preserving cost function using the perimeter of polygons created from feature points. The spatially-varying warps reduce the parallax artifact of image stitching by a certain amount, however they cannot reflect abrupt depth changes in a captured scene completely since the neighboring cells are processed with smoothness constraints. Moreover, the spatially-varying warps were inherently designed to deform images assuming small baselines [32], and thus the warped images look unnatural when the relative orders of control points are changed across multiple images due to large parallax [33].

The stitched images usually exhibit perspective distortions in non-overlapping regions among multiple images where no valid feature matches are obtained. To alleviate the

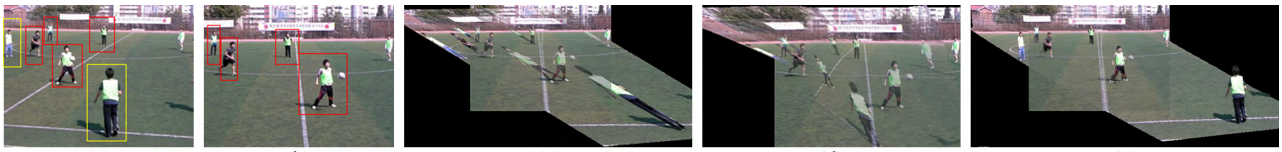


FIGURE 1. Stitching images with large parallax. (a) A target image and (b) a reference image. The resulting stitched images by using (c) a homography based warping scheme, (d) APAP [13], and (e) the proposed parallax-adaptive stitching, respectively.

perspective distortions, shape-preserving warps were proposed which extrapolate the warping models to non-overlapping regions using similarity transformation and/or homography linearization [15]–[18]. Chang *et al.* [15] applied a homography to the overlapping region of images and similarity transformations to the non-overlapping regions, respectively. Lin *et al.* [16] proposed a homography linearization method to combine homography and similarity transformations smoothly. Chen and Chuang [17] improved the shape-preserving warp by accurately estimating the scale and rotation of similarity transformation. Li *et al.* [18] proposed quasi-homography warps which linearly extrapolate the horizontal component of homography. The shape-preserving warps provide visually plausible stitching results, but do not always produce geometrically correct results.

Attempts have been also made to align only a certain region of input images and hide the artifacts of mismatched regions by applying seam-based composition methods. Gao *et al.* [19] obtained multiple homographies by taking the groups of inlier feature matches in order, and selected the best homography that yields a minimum seam cost. Zhang and Liu clustered closely located feature points together and found an optimal local homography associated with a minimum seam cutting error to align a local image region [20]. They also applied content-preserving warping (CPW) [34] to further refine the local alignment. Lin *et al.* [21] generated multiple local homographies using a superpixel-based grouping scheme, and further refined each homography to select the best one by using energy minimization. They also designed an energy function to encourage the warp undergoes similarity transformation and to preserve the structures like curves and lines after warping. Note that these techniques register one local region only and thus inevitably cause geometrically inaccurate stitching results.

On the other hand, the previous video stitching algorithms simply apply the existing image stitching techniques to stitch the video frames at each time instance, respectively [28]. Also, they extend the image stitching techniques straightforwardly to video stitching for the purposes of improving the computation speed or reducing the flickering artifacts. El-Saban *et al.* [28] computed SIFT descriptors for selected frames only and tracked the feature points to reduce the computational complexity of video stitching. Jiang and Gu extended CPW of local alignment and image composition to video stitching by applying the seam cutting scheme to spatiotemporal domain [29].

B. STATIC MULTI-CAMERA BASED TRACKING

Multi-camera based people tracking techniques detect walking pedestrians on a ground plane from multiple videos, which are captured by different static cameras set toward a common ground plane and positioned with relatively wide baselines. Specifically, moving foreground objects are first detected by background subtraction methods, and then the elongated shapes of detected people are represented by principal axes [24] which are used for people tracking in addition to the ground plane homography. To localize each person for robust tracking, Khan and Shah [25] computed multiple homographies associated with parallel planes to the ground plane using vanishing points. In addition to homography and vanishing points, fundamental matrix was also used to reliably find correspondence matching for the top points of people [26].

III. PARALLAX-ADAPTIVE PIXEL WARPING MODEL

In many practical applications of multi-view videos such as surveillance and sports, static multiple cameras are located with wide baselines toward a target real-world scene which yields severely different camera parameters, e.g., rotation, translation, and zoom factor. Also, in a typical video sequence, the background is composed of a ground plane and optionally a far distant region orthogonal to the ground plane, and moreover, people moving on the ground plane at different distances from the cameras are captured as multiple foreground objects. Figs. 1(a) and (b) show two frames of the 'Soccer' sequence captured by two cameras with severely different positions and viewing directions from each other, where large parallax is observed especially in the vicinity of the foreground objects. For example, the players denoted by red boxes in Fig. 1(a) appear in a different order in Fig. 1(b). In addition, the players denoted by yellow boxes appear in only one view of Fig. 1(a) not in Fig. 1(b).

Such large parallax makes the multi-view video stitching quite a challenging problem, and the conventional stitching techniques often fail to provide faithful results. Fig. 1(c) shows the stitched image by warping a target frame in Fig. 1(a) to a reference frame in Fig. 1(b) according to the homography. Since the homography-based warping assumes a planar scene structure, only the ground plane is accurately aligned and the foreground objects and the distant background region yield large parallax artifacts. Also, Fig. 1(d) shows the stitching result of APAP [13] which is one of the state-of-the-art image stitching techniques.

The APAP adaptively warps images using mesh grid structure to reduce parallax artifacts, however, it still exhibits inaccurate alignment of multiple foreground objects due to depth discontinuity, and furthermore, it causes perspective distortions in the non-overlapping area between two images.

The parallax between two views can be explained based on the epipolar geometry as shown in Fig. 2. Homography is a planar mapping from one image domain to another image domain. Suppose that a 3D real-world point $\mathbf{X}_1$ is located on a plane π and projected to the pixels $\mathbf{p}_1$ and $\mathbf{q}_1$ in the image planes I and J, respectively. Then the relation between $\mathbf{p}_1$ and $\mathbf{q}_1$ is described by

$$\mathbf{q}_1 = \mathbf{H}_\pi \mathbf{p}_1 \tag{1}$$

where $\mathbf{H}_\pi$ is the homography associated with the plane π. However, for the pixels $\mathbf{p}_2$ and $\mathbf{q}_2$ projected from a 3D point $\mathbf{X}_2$, which is not on π, the relation (1) does not hold, i.e., $\mathbf{q}_2 \neq \mathbf{H}_\pi \mathbf{p}_2$, and therefore, a single homography $\mathbf{H}_\pi$ map $\mathbf{p}_2$ to a wrong pixel $\tilde{\mathbf{q}}_2 = \mathbf{H}_\pi \mathbf{p}_2$, which causes parallax artifact. On the other hand, we can describe the geometric relationship between any pair of corresponding pixels by epipolar constraint. For example, for a given pixel $\mathbf{p}_2 \in I$, the corresponding pixel $\mathbf{q}_2 \in J$ should be located on the epipolar line $\mathbf{l}_2$ computed as

$$\mathbf{l}_2 = \mathbf{F}\mathbf{p}_2 \tag{2}$$

where $\mathbf{F}$ is the fundamental matrix.

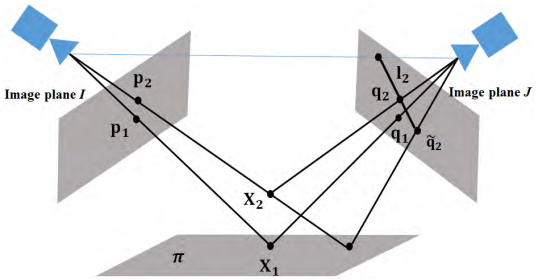


FIGURE 2. Epipolar geometry.

In this work, we propose an adaptive pixel warping model for parallax-free stitching of MVLP which employs faithful correspondence matching among multi-view videos based on the epipolar constraint. We first define on-plane pixels which are projected from the ground plane in real-world scene, and define off-plane pixels belonging to the foreground objects and the far distant background region. We generalize the concept of epipolar constraint, used for matching the top points of people in multi-camera based tracking [26], to find reliable correspondence matching of off-plane pixels. As shown in Fig. 3, for a given off-plane pixel $\mathbf{p}$ in a target image I, we first estimate the ground plane pixel (GPP) $\mathbf{g}_\mathbf{p}$ of $\mathbf{p}$ along the object direction $\mathbf{L}_\mathbf{p} = \mathbf{p} \times \mathbf{v}_\mathcal{I}$ determined by the vertical vanishing point $\mathbf{v}_\mathcal{I}$. Since $\mathbf{g}_\mathbf{p}$ is an on-plane pixel, it can be warped to the corresponding GPP $\mathbf{g}_\mathbf{q}$ in the reference

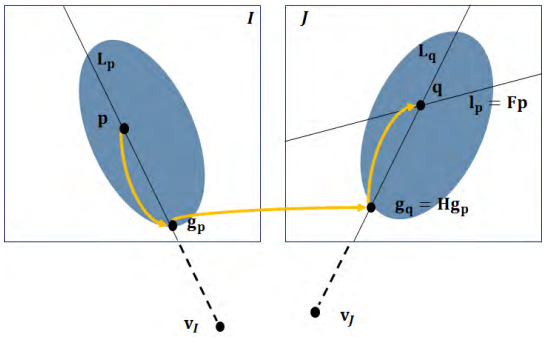


FIGURE 3. Parallax-adaptive pixel warping.

image J by using the homography matrix $\mathbf{H}$ evaluated on the ground plane.

$$\mathbf{g}_\mathbf{q} = \mathbf{H}\mathbf{g}_\mathbf{p}. \tag{3}$$

The unknown pixel $\mathbf{q}$ corresponding to $\mathbf{p}$ can be estimated as the cross point between the object direction line $\mathbf{L}_\mathbf{q} = \mathbf{g}_\mathbf{q} \times \mathbf{v}_\mathcal{J}$ passing through $\mathbf{g}_\mathbf{q}$ and the vertical vanishing point $\mathbf{v}_\mathcal{J}$, and the epipolar line $\mathbf{l}_\mathbf{p} = \mathbf{F}\mathbf{p}$ specified by the fundamental matrix.

$$\mathbf{q} = \mathbf{L}_\mathbf{q} \times \mathbf{l}_\mathbf{p}. \tag{4}$$

Fig. 1(e) shows the resulting image stitched by using the proposed warping model, where we see that the multiple foreground objects and the background are aligned correctly, while the parallax artifacts, occurred in the conventional methods as shown in Figs. 1(c) and (d), are alleviated effectively. Also the proposed algorithm can warp the foreground objects and the background on the non-overlapped areas naturally as well.

Consequently, to perform the proposed parallax-adaptive pixel warping, we need to estimate the parameters of the homography matrix $\mathbf{H}$ of the ground plane, the fundamental matrix $\mathbf{F}$, and the vertical vanishing points $\mathbf{v}_\mathcal{I}$ and $\mathbf{v}_\mathcal{J}$. We will explain the details of the parameter estimation in Section IV. Also, we need to estimate an optimal GPP $\mathbf{g}_\mathbf{p}$ for a given query pixel $\mathbf{p}$. Note that [26] employs only a single query pixel at the top of a foreground object and roughly estimates the GPP by using the average height of objects. In this work, we estimate optimal GPPs more accurately by using the spatial and temporal feature matches based on an energy minimization framework, which will be explained in Section V.

IV. PARAMETER ESTIMATION

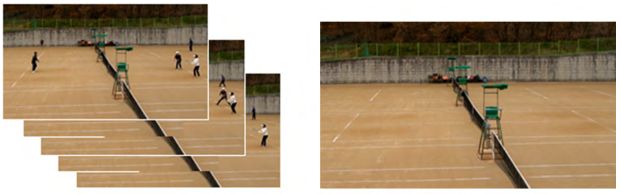
For given two input MVLP, we first estimate the parameters of the homography matrix, the fundamental matrix, and the vertical vanishing points. Note that these parameters are fixed over all the frames since we assume that multi-view videos are captured by static cameras.

A. GROUND PLANE HOMOGRAPHY

We estimate the homography associated with the ground plane using inter-view correspondence matching. In general,

initial matching between two views is performed by using feature descriptors such as SIFT [23] or ASIFT [35], and then the spurious matches are removed by outlier removal schemes such as RANSAC [36]. However, the conventional appearance-based techniques may not provide reliable matching results on MVLP, especially in multiple foreground objects at different scene depths, since the neighboring pixels of a feature point in one image yield severely different values from that of the corresponding feature point in another image [37], [38]. Therefore, in this work, we estimate the homography more reliably by employing the appearance features as well as the activity information of moving foreground objects.

Fig. 4(a) shows an input color video sequence: $\mathcal{I} = \{I^{(k)} : k = 1, 2, \cdots, K\}$ where $I^{(k)}$ denotes the k-th frame and K is the total number of frames. We find $\mathcal{B}_{\text{ground}}$ the set of feature matches on the ground plane between $\mathcal{I}$ and $\mathcal{J}$ using the activity-based correspondence matching technique [38]. Then we compute an initial homography $\mathbf{H}_{\text{init}}$ from $\mathcal{B}_{\text{ground}}$ using RANSAC. We also obtain a background image I_{BG}, as shown in Fig. 4(b), by performing the median filtering to all the frames in $\mathcal{I}$. Then we use SIFT to find a set of feature matches $\mathcal{B}$ between two background images I_{BG} and J_{BG} obtained from two video sequences $\mathcal{I}$ and $\mathcal{J}$, respectively. Note that $\mathcal{B}$ includes the matches on the ground plane and the matches in the distant background region together. Hence we first extract the matches on the ground plane only from $\mathcal{B}$ by selecting the inliers matches of $\mathbf{H}_{\text{init}}$. Then we refine $\mathbf{H}_{\text{init}}$ to obtain a final homography $\mathbf{H}$ by using $\mathcal{B}_{\text{ground}}$ and the selected ground plane matches in $\mathcal{B}$, based on RANSAC.



(a) (b)

FIGURE 4. (a) An input video sequence and (b) its background image.

B. FUNDAMENTAL MATRIX

To estimate the fundamental matrix between two views, we find inter-view feature matching on the foreground objects as well. Note that, while the correspondence matching for the background is performed once over a whole video sequence, that for the foreground objects is performed at each time instance, respectively. In practice, we use SIFT to find the inter-view feature matches between $I^{(k)}$ and $J^{(k)}$, and obtain the set $\mathcal{F}^{(k)}_{\text{spatial}}$ by selecting the matches lying on the foreground regions only by using background subtraction [39]. While $\mathcal{B}_{\text{ground}}$ includes a small number of outlier matches thanks to reliable performance of activity-based matching, $\mathcal{F}^{(k)}_{\text{spatial}}$ and $\mathcal{B}$ include relatively large numbers of spurious matches since appearance-based matching is vulnerable to severe parallax. Therefore, we further refine the matches in $\mathcal{F}^{(k)}_{\text{spatial}}$ and $\mathcal{B}$ using the geometric constraints.

As shown in Fig. 5, when a pair of corresponding off-plane pixels $\mathbf{p} \in I$ and $\mathbf{q} \in J$ are given, their GPPs $\mathbf{g_p}$ and $\mathbf{g_q}$ are corresponding on-plane pixels to each other and should be located on the object direction lines $\mathbf{L_p}$ and $\mathbf{L_q}$, respectively. Hence we can estimate $\mathbf{g_p}$ and $\mathbf{g_q}$ as [24]

$$\begin{aligned} \mathbf{g_p} &= \mathbf{L_p} \times \mathbf{L}'_{\mathbf{q}}, \\ \mathbf{g_q} &= \mathbf{L_q} \times \mathbf{L}'_{\mathbf{p}}, \end{aligned} \tag{5}$$

where $\mathbf{L}'_{\mathbf{p}}$ and $\mathbf{L}'_{\mathbf{q}}$ are the warped lines of $\mathbf{L_p}$ and $\mathbf{L_q}$ into the other views, respectively, by the ground plane homography $\mathbf{H}$. Based on this property, we induce two geometric constraints to validate the obtained correspondence matches. First, $\mathbf{g_p}$ should be located at a position on $\mathbf{L_p}$ equal to or below $\mathbf{p}$ such that $(\mathbf{g_p} - \mathbf{p}) \cdot \mathbf{v}_{\mathcal{I}} \geq 0$. Similarly, we have $(\mathbf{g_q} - \mathbf{q}) \cdot \mathbf{v}_{\mathcal{J}} \geq 0$. Second, $\mathbf{g_p}$ should be close to the lowest possible pixel $\mathbf{p}_{\text{low}}$ along $\mathbf{L_p}$ in a connected object area. In practice, we employ a tolerance range for $\mathbf{g_p}$ such that $|(\mathbf{g_p} - \mathbf{p}_{\text{low}}) \cdot \frac{\mathbf{v}_{\mathcal{I}}}{||\mathbf{v}_{\mathcal{I}}||}|$ is less than 40% of the height of a foreground object. This also applies to $\mathbf{g_q}$ and $\mathbf{q}$.

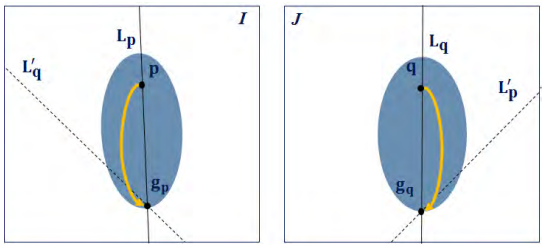


FIGURE 5. Ground plane pixel estimation. $\mathbf{p}$ and $\mathbf{q}$ are given as corresponding to each other. $\mathbf{L}'_{\mathbf{p}}$ and $\mathbf{L}'_{\mathbf{q}}$ denote the homography transformed lines of $\mathbf{L_p}$ and $\mathbf{L_q}$ into the other views, respectively.

We remove the false matches from $\mathcal{F}^{(k)}_{\text{spatial}}$, which violate the first and/or second constraints, to yield a refined set $\tilde{\mathcal{F}}^{(k)}_{\text{spatial}}$. For $\mathcal{B}$, we test only the first constraint and apply the multi-structure guided sampling (MULTI-GS) [40] to obtain a refined set $\tilde{\mathcal{B}}$. Fig. 6 shows that the proposed matching refinement for MVLP removes most of the spurious matches successfully both on the foreground objects and the background. Finally, we estimate the fundamental matrix $\mathbf{F}$ by applying RANSAC to the appearance-based feature matches of $\tilde{\mathcal{F}}^{(k)}_{\text{spatial}}$'s and $\tilde{\mathcal{B}}$ as well as the activity-based matches of $\mathcal{B}_{\text{ground}}$ together. Note that, due to computational complexity, we empirically collect 1000 feature matches from $\tilde{\mathcal{F}}^{(k)}_{\text{spatial}}$'s associated with randomly selected frames.

C. VERTICAL VANISHING POINTS

Vanishing points are the points where the parallel lines are converging [31]. In multi-view video sequences, people are assumed to be standing along the orthogonal direction to the ground plane, and therefore, we define a vertical vanishing

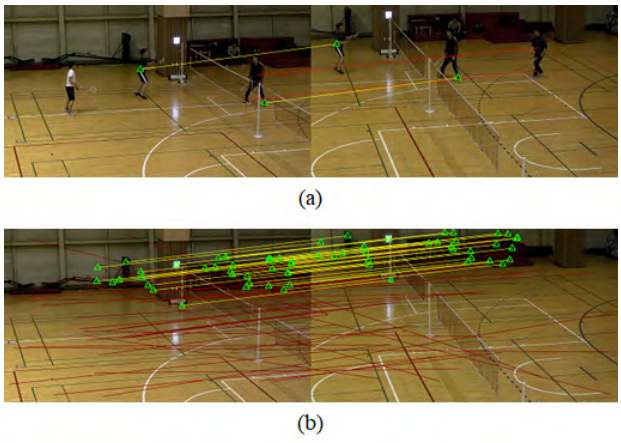


FIGURE 6. Refinement of feature matching on (a) foreground objects and (b) background. Correct and spurious matches are denoted by the yellow and red lines, respectively.

point as a converging point of parallel lines in a scene orthogonal to the ground plane. In practice, we estimate the vertical vanishing points by using [41]. Instead of complex people tracking, we simply select 10,000 major axis lines of people from randomly selected frames, where the lines satisfy the condition that the ratio of the length of minor axis to the length of major axis is below 0.3. Then, as shown in Fig. 3, the object direction $\mathbf{L_p}$ can be computed at each off-plane pixel $\mathbf{p}$ as the line passing through $\mathbf{p}$ and $\mathbf{v}$

$$\mathbf{L_p} = \mathbf{p} \times \mathbf{v} \tag{6}$$

where $\mathbf{v}$ is the vertical vanishing point. Note that the object direction $\mathbf{L_p}$ is used to estimate the GPP $\mathbf{g_p}$ based on the constraint that $\mathbf{g_p}$ should be located on $\mathbf{L_p}$.

V. GROUND PLANE PIXEL ESTIMATION

We estimate optimal GPPs for given query pixels in a target frame to find their warped pixels in a reference frame. Note that the proposed pixel warping model is not only applicable to off-plane pixels but on-plane pixels such that $\mathbf{g_p} = \mathbf{p}$ for a pixel $\mathbf{p}$ on the ground plane. We perform the GPP estimation for the foreground objects and the background, respectively, where the inter-view and inter-frame feature matches are used together for the foreground objects while only the inter-view feature matches are used for the background. The estimated GPP positions are also optimized based on an energy minimization framework.

A. GROUND PLANE PIXEL AND GROUND VALUE

Multiple off-plane pixels on a same object direction line share a same GPP, since the corresponding real-world points are assumed to be located on a same vertical line perpendicular to the ground plane. For example, as shown in Fig. 7(a), the pixels $\mathbf{r}_1$, $\mathbf{r}_2$ and $\mathbf{r}_3$ on $\mathbf{L_r}$ have the GPP $\mathbf{g_r}$, while the pixels $\mathbf{s}_1$, $\mathbf{s}_2$ and $\mathbf{s}_3$ on $\mathbf{L_s}$ have GPP $\mathbf{g_s}$. However, off-plane pixels lying on different object direction lines have different GPPs. We define a ground value $\delta_\mathbf{p}$ for the pixel $\mathbf{p}$ according to its GPP $\mathbf{g_p}$, as shown in Fig. 7(a).

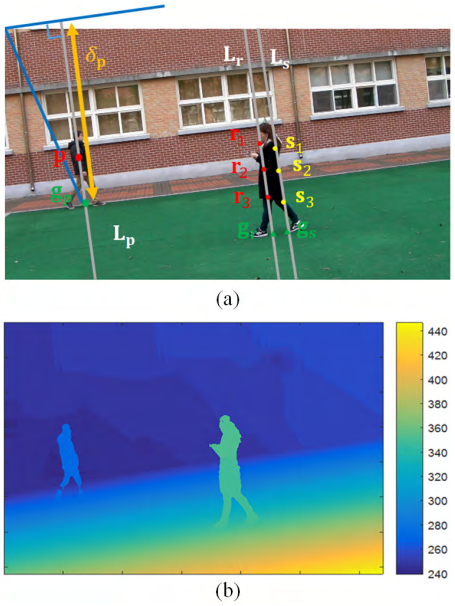


FIGURE 7. Relation between ground plane pixels and ground values. (a) A target frame and (b) its ground value map.

$$\delta_\mathbf{p} = \frac{\mathbf{v}_\mathcal{I} - \mathbf{p}}{||\mathbf{v}_\mathcal{I} - \mathbf{p}||} \cdot \mathbf{g_p}. \tag{7}$$

Note that the ground values of off-plane pixels are almost invariant within a same foreground object or a same distant background region. We exploit this property to estimate the GPPs by estimating their ground values instead, since $\mathbf{g_p}$ and $\delta_\mathbf{p}$ are put in one-to-one correspondence with each other for a given $\mathbf{p}$ via (7).

B. SPATIOTEMPORAL ESTIMATION FOR FOREGROUND OBJECTS

Let us first define $\Phi^{(k)}_{\text{spatial}}$ as the set of feature pixels in $\tilde{\mathcal{F}}^{(k)}_{\text{spatial}}$ detected from a target image $I^{(k)}$. For a given feature pixel $\mathbf{p}^{(k)} \in \Phi^{(k)}_{\text{spatial}}$ associated with an inter-view match denoted by a yellow line in Fig. 8, a GPP $\mathbf{g}_{\mathbf{p}^{(k)}}$ is found by (5). We call this procedure of GPP estimation using inter-view feature matches as spatial matching based estimation (SME). We perform SME using $\tilde{\mathcal{F}}^{(k)}_{\text{spatial}}$ for each k-th frame, respectively.

However, some foreground objects may not provide sufficient numbers of inter-view matches or may have no inter-view match at all, due to large parallax between two views and/or relatively small areas in an image. Hence we additionally employ the temporal information from the previous frame to predict GPPs. Specifically, we use SIFT to obtain the set of inter-frame feature matches $\tilde{\mathcal{F}}^{(k)}_{\text{temporal}}$ associated with the foreground objects between a current frame $I^{(k)}$ and its previous frame $I^{(k-1)}$, which are denoted by the blue lines in Fig. 8. In general, $\tilde{\mathcal{F}}^{(k)}_{\text{temporal}}$ has a much larger number of reliable matches than $\tilde{\mathcal{F}}^{(k)}_{\text{spatial}}$, since the adjacent frames in a same view exhibit similar scene contents to each other while

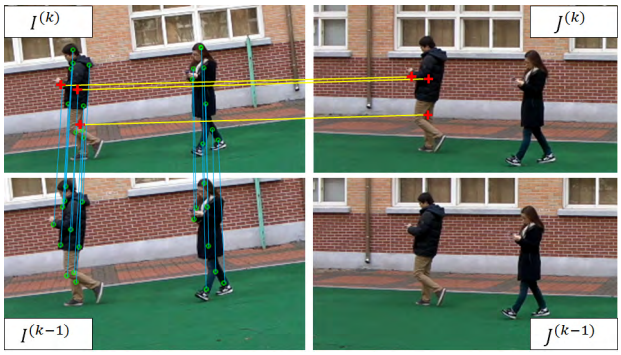


FIGURE 8. Inter-view feature matches (yellow lines) and inter-frame feature matches (blue lines).

the frames from different views exhibit severely different appearance due to large parallax. Note that some pixels may be detected as the spatial features and the temporal features simultaneously, which belong to both of $\tilde{\mathcal{F}}^{(k)}_{\text{spatial}}$ and $\tilde{\mathcal{F}}^{(k)}_{\text{temporal}}$.

Let us define $\Phi^{(k)}_{\text{temporal}}$ as the set of inter-frame feature pixels in $\tilde{\mathcal{F}}^{(k)}_{\text{temporal}}$ detected from a target image $I^{(k)}$. For each pixel $\mathbf{p}^{(k)} \in \left(\Phi^{(k)}_{\text{temporal}} - \Phi^{(k)}_{\text{spatial}}\right)$, we find its temporal corresponding pixel $\mathbf{p}^{(k-1)}$. In addition, we also collect the inter-view feature pixels from $\Phi^{(k)}_{\text{spatial}}$, which are located in the same foreground object to $\mathbf{p}^{(k)}$. Then, by (3) and (4), we compute a candidate pixel $\hat{\mathbf{q}}^{(k)}$ in the reference image $J^{(k)}$ corresponding to $\mathbf{p}^{(k)}$ by finding a candidate GPP $\hat{\mathbf{g}}_{\mathbf{p}^{(k)}}$. Note that we estimate the optimal GPP by estimating the ground value via (7) instead. In practice, we take a ground value of $\mathbf{p}^{(k)}$ as the ground value of $\mathbf{p}^{(k-1)}$ and the ground values of the additionally collected inter-view feature pixels, respectively, since the ground values are same within a same foreground object while the GPPs are changeable. Then we check whether each of the candidate positions $\hat{\mathbf{q}}^{(k)}$ lies on a foreground object region in $J^{(k)}$ or not, and we discard the associated GPP $\hat{\mathbf{g}}_{\mathbf{p}^{(k)}}$ when $\hat{\mathbf{q}}^{(k)}$ lies outside of the foreground areas within $J^{(k)}$. Finally, we evaluate the SIFT descriptors for the surviving candidate positions $\hat{\mathbf{q}}^{(k)}$, and select the GPP of $\mathbf{p}^{(k)}$ associated with the best matching candidate position. We call this procedure as temporal matching based estimation (TME).

When TME returns no available solution, we estimate the GPP by taking the ground value of the lowest possible pixel in a foreground object. We call this procedure as region based estimation (RE). RE yields relatively lower accuracy of GPP estimation than SME due to the lack of inter-view matching information, however it can perform reasonable warping of the foreground objects lying on the non-overlapping region which appear only in $I^{(k)}$ but not in $J^{(k)}$.

C. SPATIAL ESTIMATION FOR BACKGROUND

We also estimate the GPPs for the background. We assume that the background is composed of the ground plane and optionally a far distant region. To adaptively warp the background image, we first decide whether the captured scene includes a distant background region or not. To be specific, we use the inter-view feature matching on the background. From $\tilde{\mathcal{B}}$, we extract the set of matches which are outliers of the ground plane homography obtained in Section IV-A. If the number of outlier matches is less than 5% of the total number of matches in $\tilde{\mathcal{B}}$, we decide the background scene includes only the ground plane without a distant region, and then we simply estimate the GPPs as $\mathbf{g}_{\mathbf{p}} = \mathbf{p}$ for all the background pixels $\mathbf{p}$.

Otherwise, it means that the background includes a distant region where we perform GPP estimation. We first compute the GPPs for the extracted outlier matches in $\tilde{\mathcal{B}}$ by SME, and predict a line passing through the obtained GPPs using linear regression. This line is regarded as a boundary to roughly separate the distant background region from the ground plane. For the pixels $\mathbf{p}$ located below the boundary line, we simply estimate the GPPs as $\mathbf{g}_{\mathbf{p}} = \mathbf{p}$. For the feature pixels in $\tilde{\mathcal{B}}$ located above the boundary line, we estimate the GPPs by SME.

D. GROUND VALUE OPTIMIZATION

For seamless warping of foreground objects and distant background region, we further refine the positions of the initial GPPs for the off-plane feature pixels, obtained in Section V-B and Section V-C. Specifically, we formulate an energy function E_{FG} to refine the associated initial ground values for the feature pixels of the foreground objects in $\Phi^{(k)} = \left(\Phi^{(k)}_{\text{spatial}} \bigcup \Phi^{(k)}_{\text{temporal}}\right)$.

$$E_{\text{FG}}(\mathbf{F}^{(k)}) = E_{\text{FG,data}}(\mathbf{F}^{(k)}) + \alpha E_{\text{FG,ss}}(\mathbf{F}^{(k)}) + \beta E_{\text{ts}}(\mathbf{F}^{(k)}) \tag{8}$$

where $\mathbf{F}^{(k)}$ denotes the set of optimal ground values $\delta_{\mathbf{p}^{(k)}}$'s for all feature pixels $\mathbf{p}^{(k)}$'s in $\Phi^{(k)}$. We set the weighting parameters as $\alpha = 0.5$ and $\beta = 0.5$ experimentally. $E_{\text{FG,data}}$ is the data cost designed as

$$E_{\text{FG,data}}(\mathbf{F}^{(k)}) = \sum_{\mathbf{p}^{(k)} \in \Phi^{(k)}} \left(\delta_{\mathbf{p}^{(k)}} - \bar{\delta}_{\mathbf{p}^{(k)}}\right)^2 \tag{9}$$

where $\bar{\delta}_{\mathbf{p}^{(k)}}$ denotes the initial ground value of $\mathbf{p}^{(k)}$. The initial ground values may be inaccurate due to the errors in feature matching and/or background subtraction. Hence we employ the spatial smoothness cost given by

$$E_{\text{FG,ss}}(\mathbf{F}^{(k)}) = \sum_{\mathbf{p}_i^{(k)} \in \Phi^{(k)}} \sum_{\mathbf{p}_j^{(k)} \in \mathcal{N}_i^{(k)}} w(\mathbf{p}_i^{(k)}, \mathbf{p}_j^{(k)}) \cdot \left(\delta_{\mathbf{p}_i^{(k)}} - \delta_{\mathbf{p}_j^{(k)}}\right)^2 \tag{10}$$

where $\mathcal{N}_i^{(k)}$ denotes the set of spatially neighboring pixels to $\mathbf{p}_i^{(k)}$. Two pixels $\mathbf{p}_i^{(k)}$ and $\mathbf{p}_j^{(k)}$ are regarded as spatial neighbors to each other when they are located in a same foreground object region and satisfy the compatibility constraint: the warped pixel of $\mathbf{p}_i^{(k)} \in I^{(k)}$ using the initial GPP of $\mathbf{p}_j^{(k)}$ is located on a foreground object region in $J^{(k)}$, and at the same time, the warped pixel of $\mathbf{p}_j^{(k)} \in I^{(k)}$ using the initial GPP of $\mathbf{p}_i^{(k)}$ is also located on the same foreground object in $J^{(k)}$.

In this work, we select at most the four nearest neighboring pixels to $\mathbf{p}_i^{(k)}$ to define $\mathcal{N}_i^{(k)}$. The spatial weight is given by

$$w(\mathbf{p}_i, \mathbf{p}_j) = \exp\left(-||\mathbf{p}_i - \mathbf{p}_j||/\tau\right) \tag{11}$$

where we set $\tau = 100$ empirically. Moreover, to mitigate the flickering artifacts in a resulting stitched video sequence, the temporal smoothness cost is defined as

$$E_{\text{ts}}(\mathbf{F}^{(k)}) = \sum_{\mathbf{p}^{(k)} \in \Phi_{\text{temporal}}^{(k)}} \left(\delta_{\mathbf{p}^{(k)}} - \delta^*_{\mathbf{p}^{(k-1)}}\right)^2 \tag{12}$$

where $\delta^*_{\mathbf{p}^{(k-1)}}$ is the optimal ground value of the inter-frame corresponding pixel $\mathbf{p}^{(k-1)}$ in the previous frame $I^{(k-1)}$. Note that we do not use the temporal cost function at the first frame.

Let Ψ represent the set of the feature pixels in $\tilde{\mathcal{B}}$ located above the boundary line in the background image of a target view. We also formulate an energy function E_{BG} for Ψ as

$$E_{\text{BG}}(\mathbf{B}) = E_{\text{BG,data}}(\mathbf{B}) + \gamma E_{\text{BG,ss}}(\mathbf{B}) \tag{13}$$

where $\mathbf{B}$ denotes the set of optimal ground values $\delta_{\mathbf{p}}$'s for all feature pixels $\mathbf{p}$'s in Ψ. The weighting parameter γ is set to be 1 empirically. The data term is given by

$$E_{\text{BG,data}}(\mathbf{B}) = \sum_{\mathbf{p} \in \Psi} \left(\delta_{\mathbf{p}} - \bar{\delta}_{\mathbf{p}}\right)^2 \tag{14}$$

where $\bar{\delta}_{\mathbf{p}}$ denotes the ground value of $\mathbf{p} \in \Psi$ initially obtained by SME. The spatial smoothness cost is given by

$$E_{\text{BG,ss}}(\mathbf{B}) = \sum_{\mathbf{p}_i \in \Psi} \sum_{\mathbf{p}_j \in \mathcal{N}_i} w(\mathbf{p}_i, \mathbf{p}_j) \cdot \left(\delta_{\mathbf{p}_i} - \delta_{\mathbf{p}_j}\right)^2 \tag{15}$$

where $\mathcal{N}_i$ is the set of the four feature points in Ψ nearest to $\mathbf{p}_i$.

We refine the ground values for all the off-plane feature pixels in the foreground objects by minimizing the energy function in (8) using a linear solver. Then the remaining non-feature pixels in the foreground objects are assigned ground values by using the nearest interpolation on the available optimal ground values computed at the feature pixels. We also find the set of the optimal ground values at the off-plane feature pixels in the distant background region by minimizing the energy function in (13), which are then interpolated to determine the ground values at all the background pixels above the boundary line. In practice, we apply the linear interpolation within the convex hull of the feature pixels and apply the nearest interpolation outside of the convex hull. Fig. 7(b) shows the resulting ground value map of a target image frame in Fig. 7(a). Note that the off-plane pixels belonging to a same foreground object region or a distant background region have almost same ground values to one another, even though their GPPs are different. On the contrary, the on-plane pixels on the ground plane have different ground values according to their relative positions along the direction toward the vertical vanishing point.

VI. EXPERIMENTAL RESULTS

We evaluate the performance of the proposed algorithm using 12 test video sequences, as shown in Fig. 12. Each test video sequence is composed of two videos captured at 30 frames per second by two synchronized cameras with unknown camera parameters. A captured scene includes multiple moving people on a ground plane at various scene depths. Table 1 presents the specification of the test sequences. We simply approximate the parallax angle by first taking the sum of the angle between $\mathbf{L_p}$ and $\mathbf{L}'_{\mathbf{q}}$ and the angle between $\mathbf{L_q}$ and $\mathbf{L}'_{\mathbf{p}}$ shown in Fig. 5, and by computing the average for all the manually obtained ground truth matching pixels of $\mathbf{p}$ and $\mathbf{q}$ which is then divided by 2. In general, a larger parallax angle is yielded, when two videos are captured with a wider camera baseline and a captured scene is closer to the cameras. We warp each pixel in a target image frame to a reference frame based on the proposed parallax-adaptive pixel warping model. The hole pixels in warped target frame are interpolated by using the valid warped pixels. To evaluate whether the alignment is geometrically accurate or not, we simply use the average blending scheme to combine the warped target frame and the reference frame.

TABLE 1. Specification of test video sequences.

Sequence	Resolution	Distant Background	Time (min)	Parallax Angle(°)
Fountain	640×360	x	51	1.9
Tennis	640×360	o	34	12.3
Lawn	640×360	x	37	12.7
Badminton	640×360	o	52	18.2
Square	640×360	x	35	18.3
Office	640×360	o	30	18.5
Trail	640×360	o	51	18.7
Stadium	640×360	o	35	24.4
Soccer	320×240	o	55	28.0
Street	640×360	o	29	30.5
School	640×360	o	36	31.9
Garden	640×360	o	24	32.0

A. FOREGROUND OBJECT ALIGNMENT

The performance of video stitching highly depends on the accuracy of correspondence matching between different views. In particular, accurate inter-view matches on the foreground object regions are required to adaptively alleviate the parallax artifacts caused by different scene depths of multiple objects. Therefore, we first evaluate the alignment performance of multiple foreground objects according to various GPP estimation methods.

Fig. 9 compares the stitching results on selected frames from the three test sequences of MVLP, using the GPPs estimated by the four different methods: RE, SME+RE, SME+TME+RE without optimization, and SME+TME+RE with optimization. Figs. 9(a) and (b) show target frames and reference frames, respectively, where we mark the obtained inter-view feature pixels in $\tilde{\mathcal{F}}_{\text{spatial}}^{(k)}$ by crosses. In the "Lawn" sequence, the foreground objects

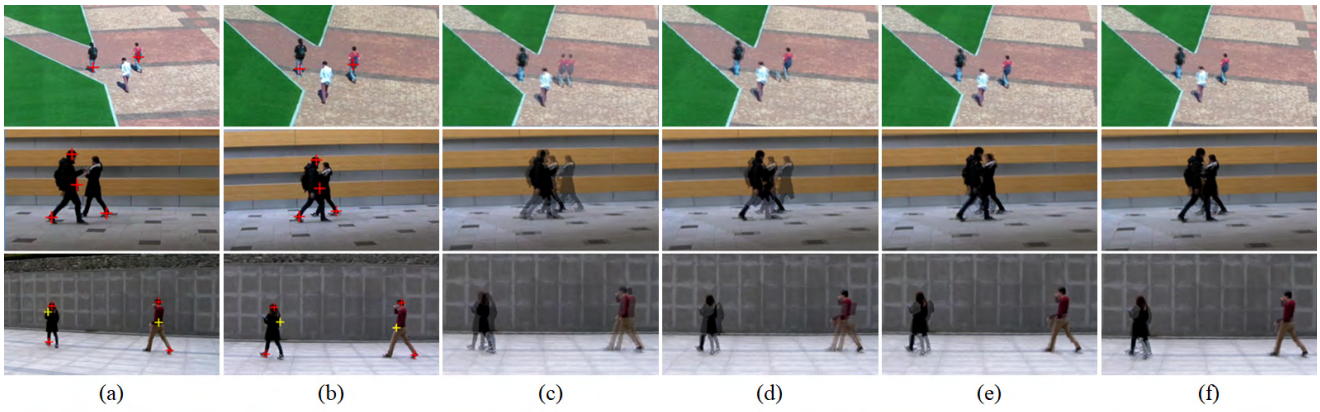


FIGURE 9. Stitching results of multiple foreground objects using the proposed ground plane pixel estimation methods. (a) Target frames and (b) reference frames. The stitched images by using (c) RE, (d) SME+RE, (e) SME+TME+RE without optimization, and (f) SME+TME+RE with optimization, respectively. From top to bottom, "Lawn," "Street," and "Garden" sequences.

occupy relatively small image areas since the cameras are located far from the captured scene, and thus they yield few inter-view feature matches. RE shows the artifact on the person in red, since the associated GPPs are selected on the person in white, which is connected to the person in red in the target frame by the blob analysis. The matching accuracy on the person in red is improved by using SME+RE, but the artifact on the legs is still observed. The selected frames in the "Street" sequence are quite a challenging case, since the two people occlude each other. SME+RE improves the results of RE using inter-view matching information, but it still causes the misalignment on the right person. However, SME+TME+RE provides accurate results of foreground object alignment on the two sequences by using the spatiotemporal information together. The "Garden" sequence includes the false matches marked by yellow crosses in Figs. 9(a) and (b). Hence, SME+RE and even SME+TME+RE without optimization suffer from the misalignment artifact of foreground objects, however, this artifact is alleviated in SME+TME+RE with optimization.

We also quantitatively measure the matching errors of the foreground objects using the ground truth correspondence matches. We select regularly distributed query pixels on the foreground objects in a target frame, and obtain initial matching pixels in a reference frame by using a dense feature descriptor DAISY [42], which are then refined manually. We find ground truth matches on 100 selected pairs of frames for each sequence, and on average, we obtain about 20 matches on the foreground objects at each pair of frames. Fig. 10 compares the root mean squared errors (RMSEs) of the foreground matching averaged over the 12 test sequences, where the RMSEs of RE, SME+RE, and SME+TME+RE without and with optimization are 5.45, 4.38, 3.77, and 3.34 pixels, respectively.

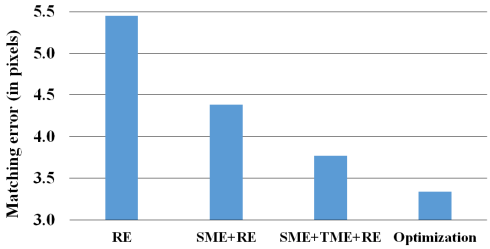


FIGURE 10. Comparison of the average error of correspondence matching for the foreground objects using different ground plane pixel estimation methods. The matching error measures the RMSE between the resulting matches and the ground truth matches averaged over 12 test sequences.

B. VIDEO STITCHING

Fig. 11 shows the video stitching results of the proposed algorithm on six test sequences of MVLP. We select frames at five different time instances in each sequence which include various challenging scene contents. In Fig. 11, all the sequences except the "Square" are detected to include the distant background regions in addition to the ground planes. We see that the ground planes and the distant background regions are well aligned simultaneously, since the on-plane pixels and the off-plane background pixels are warped adaptively. Note that the ground planes in the "Office" and "Soccer" sequences have less textures, which are often occurred in surveillance and sports scenes, but the proposed algorithm also finds correct homographies for these ground planes by using the appearance and activity based feature matches together.

We also observe that the multiple foreground objects are accurately aligned without ghosting artifacts in most frames. For example, in the "Tennis" sequence, the two people on the right side are moving toward different directions from each other, and thus they are detected as a single object at some time instances due to overlap. The proposed algorithm provides accurate warping of these foreground objects by estimating optimal GPPs reliably using the spatiotemporal feature matches. In the "Square" sequence, the left person moves on the overlapped area between the target and reference views at the 29571^{th} and 29663^{th} frames, however it disappears from the reference frames at the 29804^{th} and 29857^{th} frames. The proposed algorithm warps this object naturally on the non-overlapped area in the stitched images.

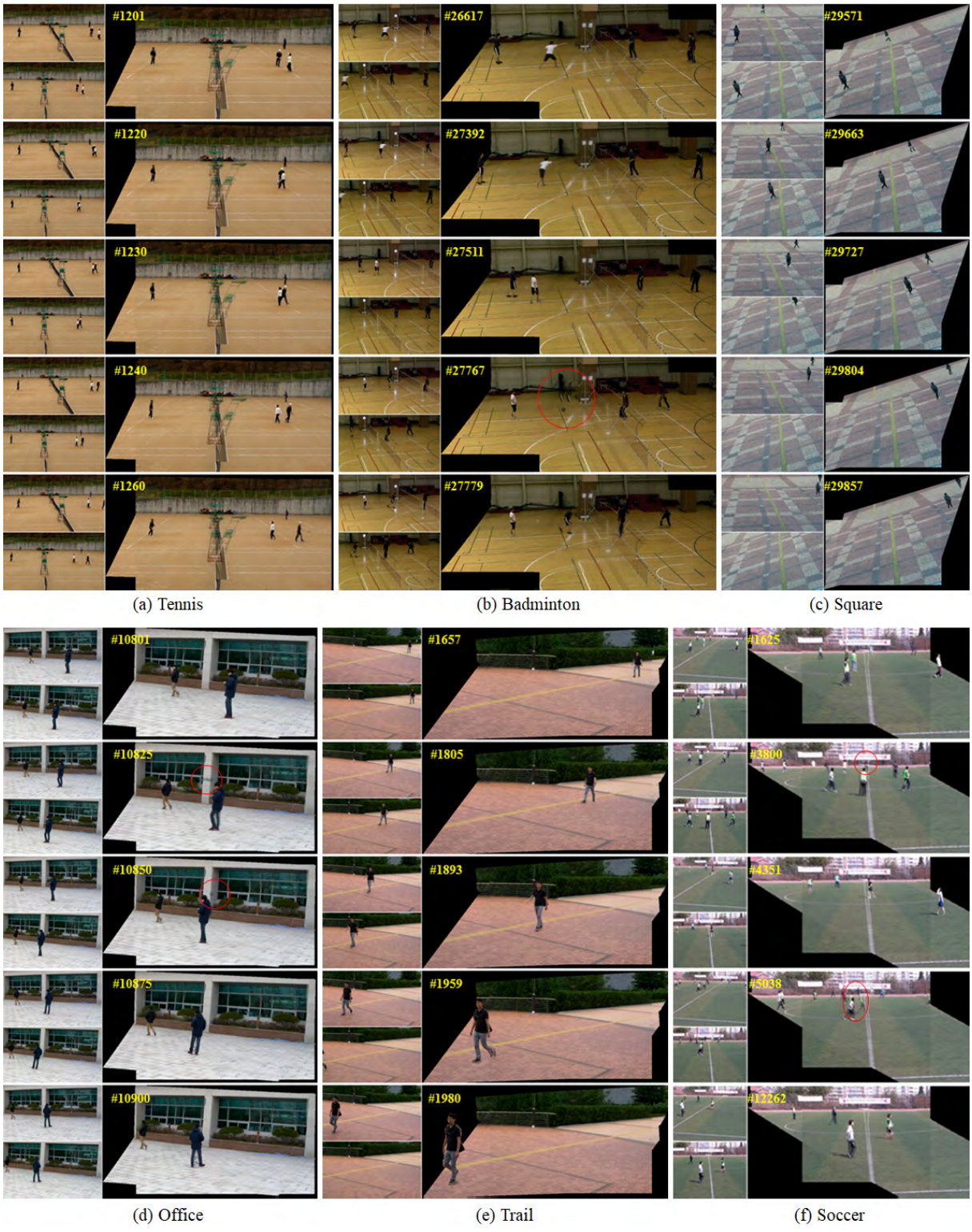


FIGURE 11. Video stitching results of the proposed algorithm. For each sequence, pairs of target and reference frames (left) and the stitched images (right) are shown.

In the ''Trail'' sequence, the foreground object approaches to the camera yielding severely changing scene depths, but the proposed algorithm aligns this object correctly at various scales. On the other hand, the proposed algorithm yields artifacts on some exceptional situations. In the ''Badminton'' sequence, the person marked with a red circle is jumping and

never touches the ground plane at the 27767th frame. In such a case, no valid inter-view feature matches are obtained on this region due to the geometric constraint in Section IV-B, and thus RE yields the misalignment artifact. In the ''Office'' sequence, we see some artifacts near the right person since a moving car behind the cameras is reflected on the background windows. The ''Soccer'' is quite a challenging sequence which includes various fast moving players, where multiple people occlude one another at the 3800th and 5038th frames. In such cases, SIFT provides insufficient correct inter-view matches or even no correct match at all, resulting in the stitching artifacts indicated by red circles.

C. COMPARISON WITH CONVENTIONAL METHODS

We compare the performance of the proposed algorithm with that of the four conventional methods including the state-of-the-art image stitching techniques: Homography, CPW [34], SPHP [15] and APAP [13]. Note that CPW is used as an alignment model for stitching methods [20], [29]. SPHP is a shape-preserving warping method which can be compared to evaluate the naturalness of warping on non-overlapping regions. APAP is one of the most flexible warping methods which directly estimates multiple homographies for local image regions. However, we do not compare the seam-based techniques [19]–[21], since they just hide the misalignment artifacts using seam-cutting based composition. We apply the compared image stitching techniques to the frames at each time instance, respectively. We implement Homography and CPW. The parameters for warp in CPW are set as [29]. We obtain the stitching results of SPHP and APAP using the source codes provided by the authors' webpages [43], [44]. In our experiment, MULTI-GS [40] used in [13] yields a better performance of outlier removal than RANSAC, and thus we also apply MULTI-GS to remove outlier matches of SIFT in Homography, CPW, and SPHP as well.

Fig. 12 compares the stitching results on selected frames of 12 test video sequences. All the conventional methods including the proposed algorithm achieve good stitching results on the ''Fountain'' sequence which yields the smallest parallax angle of 1.9°. However, for the other sequences of MVLP, the conventional methods fail to work to align multiple foreground objects and background simultaneously. For example, in the ''Square'' and ''Office'' sequences, the feet of multiple people are well aligned on the ground planes, but the mismatch artifact gets worse toward the heads, since the ground plane warping is dominant in the conventional methods. On the other hand, in the ''Stadium,'' ''Soccer,'' and ''Garden'' sequences, a same person appears twice at different locations without any overlap on the stitched domain, since the conventional methods extract dominant features from the distant background regions causing the misalignment artifacts on the ground planes and the foreground objects.

Specifically, Homography warps all the pixels in a target frame by global transformation derived from a dominant planar scene structure, and thus it mismatches either the ground plane or a distant background region. CPW adaptively refines the initial homography according to feature matches, and reduces the parallax artifacts on the foreground objects compared with that of Homography, as shown in the ''Tennis,'' ''Office'' and ''Street'' sequences. SPHP adopts the similarity transformation to reduce the perspective distortion of the non-overlapping area, and thus it aligns the foreground objects on the non-overlapping areas well in the ''Square'' sequence as marked with a red circle. However, at the same time, SPHP distorts the line structure on the ground plane to curves as marked with green ellipses in the ''Lawn'' and ''Square'' sequences. APAP estimates locally adaptive warps and reduces the spatial deviation of a same foreground object in the stitched domain compared with that of CPW, as shown in the ''School'' sequence, however, APAP results in unnatural distortions in the ''Badminton,'' ''Trail,'' and ''School'' sequences as marked with green ellipses.

On the contrary, in all the frames, the proposed algorithm alleviates the parallax artifacts of video stitching successfully by adaptively aligning the multiple foreground objects and background simultaneously. It also performs geometrically accurate warping on the non-overlapping areas as well, as shown in the ''Badminton,'' ''Square,'' and ''Soccer'' sequences. Moreover, the proposed algorithm correctly determines the existence of distant background regions in all 12 test sequences. Thus both of the ground plane and the distant background region are correctly aligned as shown in the ''Badminton,'' ''Office,'' and ''School'' sequences. In the ''Soccer'' sequence, even some ghost artifacts are observed due to significant amount of occlusion as marked by a red circle, the proposed algorithm aligns most people accurately while the compared methods fail to work on this challenging case. Also, the umpire chair and the net in the ''Tennis'' sequence and the net and the light lamp in the ''Badminton'' sequence are static objects over a whole video sequence which are not detected as moving foreground objects, and therefore the proposed algorithm cannot align them correctly. However, all the compared methods also fail to align these objects as marked with yellow ellipses. More comparative results of video stitching are provided in the supplementary video.

We also quantitatively compare the performance of the proposed algorithm with that of the conventional methods using manually obtained ground truth correspondence matches on the foreground objects and the background together. We use the same ground truth matches on the foreground objects as explained in Sec. VI-A. We generate ground truth matches on the background only once for each sequence using the background image. We first consider multiple large planar areas in the background, and compute an optimal homography for each planar area by using manually obtained feature matches. Then we select regularly distributed query pixels on the background image of a target view, and find the ground truth matching pixels by warping the query pixels employing the multiple homographies selectively. For the query pixels on small and/or non-planar areas, we manually

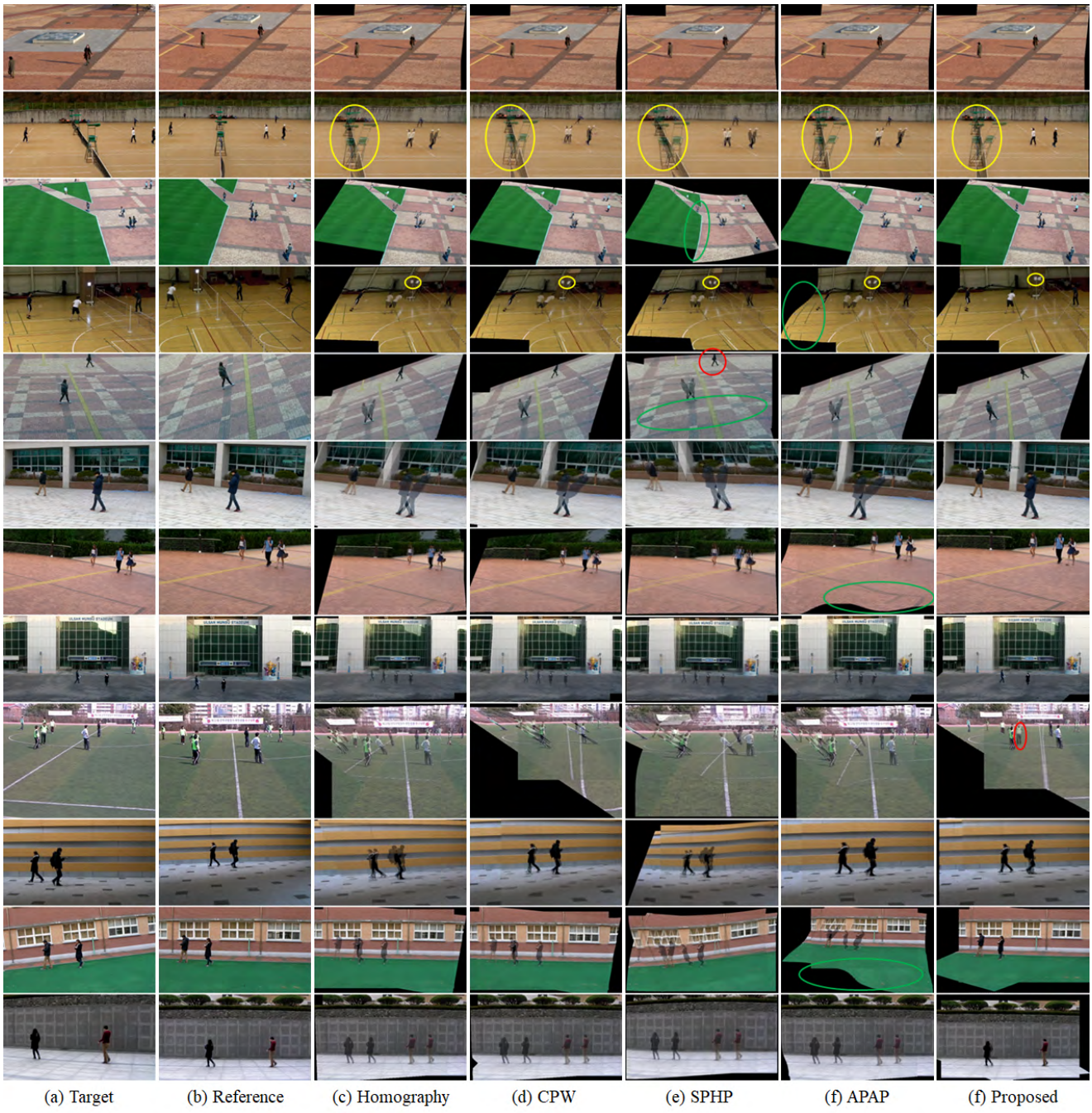


FIGURE 12. Comparison of video stitching results of the proposed algorithm and the four existing methods: Homography, CPW [34], SPHP [15], and APAP [13]. From top to bottom, "Fountain," "Tennis," "Lawn," "Badminton," "Square," "Office," "Trail," "Stadium," "Soccer," "Street," "School," and "Garden" sequences.

obtain the ground truth matching pixels. The resulting ground truth matches on the background image are added to each of the 100 frames which are selected for finding ground truth matches on the foreground objects, where we exclude the background query pixels occluded by the foreground objects. Consequently, on average, we have 724 ground truth matches on the background for each of the 100 selected frames over 12 test sequences.

Fig. 13 presents the RMSE between the ground truth corresponding pixels and the warped pixels on the overlapped regions of the target and reference frames. We see that the conventional methods tend to yield large RMSEs on test sequences with large parallax angles. For example, the RMSEs of all the stitching methods are below 2 pixels on the "Fountain" sequence which exhibits the smallest parallax angle of 1.9°. However, on the challenging sequences of

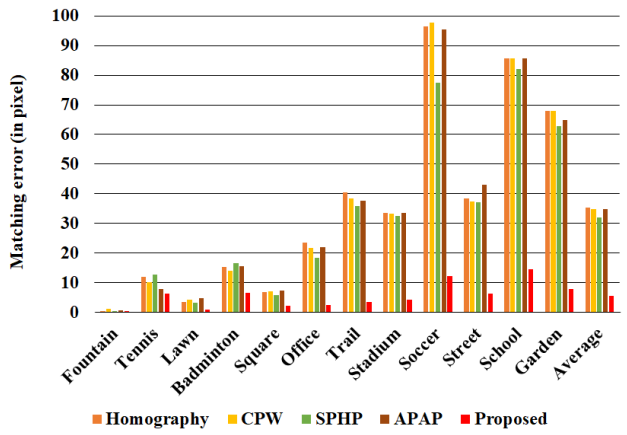


FIGURE 13. Quantitative comparison of the stitching performance of the proposed algorithm with that of the conventional methods. The matching error measures the average RMSE between the warped pixels and the ground truth corresponding pixels.

MVLP such as "Soccer" and "School," the conventional methods yield significantly larger RMSEs compared with that of the other sequences. On the other hand, the proposed algorithm always achieves smaller RMSEs than that of the conventional methods on all the test sequences, and yields a much smaller average error of 5.64 pixels while Homography, CPW, SPHP, and APAP result in the average errors of 35.37, 34.91, 32.05, and 34.86 pixels, respectively.

D. EXECUTION TIME COMPARISON

Table 2 compares the execution times of the conventional methods and the proposed algorithm measured on a PC with 3.4 GHz AMD Ryzen 7 1700X CPU and 32 GB RAM. Note that this may not be a fair comparison since the optimization level of implementation is different for the compared methods. The execution times of the conventional methods and the stitching (ST) in the proposed algorithm are averaged on 100 frames for each sequence, and that of the preprocessing (PP) and the parameter estimation (PE) in the proposed algorithm are averaged on the entire frames for each sequence. Homography is the fastest method which takes 0.57 seconds per each frame on average. CPW, SPHP, and APAP require relatively longer execution times, since these methods use different warping models for each cell or mesh grid in an image. Note that CPW is a non-parametric warping scheme and takes the longest execution time of 15.4 seconds per frame among the four conventional methods. The proposed algorithm is divided into three steps to evaluate the execution times. PP includes the background subtraction and the activity extraction for activity-based correspondence matching [38]. PE includes the homography estimation with activity-based correspondence matching computation, the fundamental matrix estimation, and the estimation of vertical vanishing points. ST includes the SIFT matching computation, ground pixel estimation, warping, and blending. Note that PP and PE are performed once over the entire frames for each video sequence, and thus yield relatively short execution times for each frame. However, ST in the proposed algorithm consumes a major portion of the execution time to compute hole pixels in the warped target frame using valid warped pixels, which takes 33.8 seconds per frame on average. Note that "Fountain," "Lawn," and "Square" sequences exhibit relatively short execution times of ST, since they do not have distant background regions.

TABLE 2. Comparison of execution times of the conventional methods and the proposed algorithm. The Unit is seconds per frame. PP: preprocessing. PE: parameter estimation. ST: stitching.

Sequence	Homography	CPW	SPHP	APAP	Proposed		
					PP	PE	ST
Fountain	0.58	13.7	5.20	4.83	0.25	0.24	9.10
Tennis	0.62	15.0	4.17	2.90	0.28	0.07	26.3
Lawn	0.54	14.8	3.91	2.75	0.27	0.14	8.20
Badminton	0.53	16.5	4.03	2.57	0.31	0.18	39.0
Square	0.59	18.1	3.91	2.55	0.29	0.17	10.0
Office	0.55	17.7	4.41	3.02	0.28	0.10	38.0
Trail	0.65	18.7	3.93	2.77	0.30	0.20	34.1
Stadium	0.69	17.2	4.20	2.87	0.30	0.08	44.8
Soccer	0.25	9.10	3.04	4.28	0.21	0.03	19.3
Street	0.57	12.2	4.19	3.32	0.30	0.09	66.0
School	0.61	16.2	4.20	3.20	0.29	0.10	39.3
Garden	0.71	15.8	4.96	4.28	0.28	0.07	72.1
Average	0.57	15.4	4.18	3.28	0.28	0.12	33.8

VII. CONCLUSIONS

We proposed a novel video stitching algorithm to achieve geometrically accurate alignment of MVLP. We warped the multiple foreground objects, distant background, and ground plane adaptively based on the epipolar geometry, where an off-plane pixel in a target view is warped to a reference view through its GPP. We also estimated optimal GPPs for the foreground objects by using the spatiotemporal feature matches, and for the background by using the spatial feature matches, respectively. The initially obtained GPPs are refined by energy minimization. Experimental results demonstrated that the proposed algorithm aligns various MVLP accurately, and yields a significantly better performance of parallax artifact reduction qualitatively and quantitatively compared with the state-of-the-art image stitching techniques. Our future research topics include the warping of static objects with large parallax and the parallax-free stitching for MVLP captured by moving cameras.

REFERENCES

[1] W. Liu, M. Zhang, Z. Luo, and Y. Cai, "An ensemble deep learning method for vehicle type classification on visual traffic surveillance sensors," *IEEE Access*, vol. 5, p. 24417–24425, 2017.
[2] R. Panda and A. K. Roy-Chowdhury, "Multi-view surveillance video summarization via joint embedding and sparse optimization," *IEEE Trans. Multimedia*, vol. 19, no. 9, pp. 2010–2021, Sep. 2017.
[3] M. Wang, B. Cheng, and C. Yuen, "Joint coding-transmission optimization for a video surveillance system with multiple cameras," *IEEE Trans. Multimedia*, vol. 20, no. 3, pp. 620–633, Mar. 2018.

[4] K. Bilal, A. Erbad, and M. Hefeeda, "Crowdsourced multi-view live video streaming using cloud computing," *IEEE Access*, vol. 5, pp. 12635–12647, 2017.
[5] S. A. Pettersen *et al.*, "Soccer video and player position dataset," in *Proc. ACM Multimedia Syst. Conf.*, 2014, pp. 18–23.
[6] Q. Yao, H. Sankoh, K. Nonaka, and S. Naito, "Automatic camera self-calibration for immersive navigation of free viewpoint sports video," in *Proc. Int. Conf. Multimedia Signal Process.*, Sep. 2016, pp. 1–6.
[7] B. Kwon, J. Kim, K. Lee, Y. K. Lee, S. Park, and S. Lee, "Implementation of a virtual training simulator based on 360° multi-view human action recognition," *IEEE Access*, vol. 5, pp. 12496–12511, 2017.
[8] B. Macchiavello, C. Dorea, E. M. Hung, G. Cheung, and W.-T. Tan, "Loss-resilient coding of texture and depth for free-viewpoint video conferencing," *IEEE Trans. Multimedia*, vol. 16, no. 3, pp. 711–725, Apr. 2014.
[9] L. Toni, G. Cheung, and P. Frossard, "In-network view synthesis for interactive multiview video systems," *IEEE Trans. Multimedia*, vol. 18, no. 5, pp. 852–864, May 2016.
[10] R. Szeliski, "Image alignment and stitching: A tutorial," *Found. Trends Comput. Graph. Vis.*, vol. 2, no. 1, pp. 1–104, 2006.
[11] J. Gao, S. J. Kim, and M. S. Brown, "Constructing image panoramas using dual-homography warping," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 49–56.
[12] W.-Y. Lin, S. Liu, Y. Matsushita, T.-T. Ng, and L.-F. Cheong, "Smoothly varying affine stitching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 345–352.
[13] J. Zaragoza, T.-J. Chin, Q.-H. Tran, M. S. Brown, and D. Suter, "As-projective-as-possible image stitching with moving DLT," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 7, pp. 1285–1298, Jul. 2014.
[14] G. Zhang, Y. He, W. Chen, J. Jia, and H. Bao, "Multi-viewpoint panorama construction with wide-baseline images," *IEEE Trans. Image Process.*, vol. 25, no. 7, pp. 3099–3111, Jul. 2016.
[15] C.-H. Chang, Y. Sato, and Y.-Y. Chuang, "Shape-preserving half-projective warps for image stitching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Sep. 2014, pp. 3254–3261.
[16] C.-C. Lin, S. U. Pankanti, K. N. Ramamurthy, and A. Y. Aravkin, "Adaptive as-natural-as-possible image stitching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1155–1163.
[17] Y.-S. Chen and Y.-Y. Chuang, "Natural image stitching with the global similarity prior," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 186–201.
[18] N. Li, Y. Xu, and C. Wang, "Quasi-homography warps in image stitching," *IEEE Trans. Multimedia*, to be published.
[19] J. Gao, Y. Li, T.-J. Chin, and M. S. Brown, "Seam-driven image stitching," in *Proc. Eurographics*, 2013, pp. 45–48.
[20] F. Zhang and F. Liu, "Parallax-tolerant image stitching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 3262–3269.
[21] K. Lin, N. Jiang, L.-F. Cheong, M. Do, and J. Lu, "SEAGULL: Seam-guided local alignment for parallax-tolerant image stitching," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 370–385.
[22] M. Yu and G. Ma, "360° surround view system with parking guidance," *SAE Int. J. Commercial Vehicles*, vol. 7, no. 1, pp. 19–24, 2014.
[23] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.
[24] W. Hu, M. Hu, X. Zhou, T. Tan, J. Luo, and S. Maybank, "Principal axis-based correspondence between multiple cameras for people tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 4, pp. 663–671, Apr. 2006.
[25] M. Shah and S. M. Khan, "Tracking multiple occluding people by localizing on multiple scene planes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 3, pp. 505–519, Mar. 2009.
[26] A. Yildiz and Y. S. Akgul, "A fast method for tracking people with multiple cameras," in *Proc. Eur. Conf. Comput. Vis. Workshops*, 2010, pp. 128–138.
[27] M. Takahashi, K. Ikeya, M. Kano, H. Ookubo, and T. Mishina, "Robust volleyball tracking system using multi-view cameras," in *Proc. Int. Conf. Pattern Recognit.*, Dec. 2016, pp. 2740–2745.
[28] M. El-Saban, M. Izz, and A. Kaheel, "Fast stitching of videos captured from freely moving devices by exploiting temporal redundancy," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2010, pp. 1193–1196.
[29] W. Jiang and J. Gu, "Video stitching with spatial-temporal content-preserving warping," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2015, pp. 42–48.
[30] K.-Y. Lee and J.-Y. Sim, "Robust video stitching using adaptive pixel transfer," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2015, pp. 813–817.
[31] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*. Cambridge, U.K.: Cambridge Univ. Press, 2003.
[32] T. Igarashi, T. Moscovich, and J. F. Hughes, "As-rigid-as-possible shape manipulation," *ACM Trans. Graph.*, vol. 24, no. 3, pp. 1134–1141, Jul. 2005.
[33] S. Schaefer, T. McPhail, and J. Warren, "Image deformation using moving least squares," *ACM Trans. Graph.*, vol. 25, no. 3, pp. 533–540, 2006.
[34] F. Liu, M. Gleicher, H.-L. Jin, and A. Agarwala, "Content-preserving warps for 3D video stabilization," *ACM Trans. Graph.*, vol. 28, no. 3, pp. 44:1–44:9, Aug. 2009.
[35] J.-M. Morel and G. Yu, "ASIFT: A new framework for fully affine invariant image comparison," *SIAM J. Imag. Sci.*, vol. 2, no. 2, pp. 438–469, 2009.
[36] M. A. Fischler and R. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM*, vol. 24, no. 6, pp. 381–395, 1981.
[37] E. Ermis, P. Clarot, P. Jodoin, and V. Saligrama, "Activity based matching in distributed camera networks," *IEEE Trans. Image Process.*, vol. 19, no. 10, pp. 2595–2613, Oct. 2010.
[38] S.-Y. Lee, J.-Y. Sim, C.-S. Kim, and S.-U. Lee, "Correspondence matching of multi-view video sequences using mutual information based similarity measure," *IEEE Trans. Multimedia*, vol. 15, no. 8, pp. 1719–1731, Dec. 2013.
[39] J. M. McHugh, J. Konrad, V. Saligrama, and P.-M. Jodoin, "Foreground-adaptive background subtraction," *IEEE Signal Process. Lett.*, vol. 16, no. 5, pp. 390–393, May 2009.
[40] T.-J. Chin, J. Yu, and D. Suter, "Accelerated hypothesis generation for multistructure data via preference analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 4, pp. 625–638, Apr. 2012.
[41] F. Lv, T. Zhao, and R. Nevatia, "Camera calibration from video of a walking human," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 9, pp. 1513–1518, Sep. 2006.
[42] E. Tola, V. Lepetit, and P. Fua, "DAISY: An efficient dense descriptor applied to wide-baseline stereo," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 5, pp. 815–830, May 2010.
[43] *C.-H. Chang's Page*. Accessed: May 12, 2018. [Online]. Available: https://www.cmlab.csie.ntu.edu.tw/~frank/
[44] *Project Page of APAP*. Accessed: May 12, 2018. [Online]. Available: http://cs.adelaide.edu.au/~tjchin/apap/

KYU-YUL LEE received the B.S. degree in electrical and computer engineering from the Ulsan National Institute of Science and Technology, Ulsan, South Korea, in 2013, where he is currently pursuing the Ph.D. degree in electrical and computer engineering. His research interests include correspondence matching, video stitching, and deep learning.

JAE-YOUNG SIM (S'02–M'06) received the B.S. degree in electrical engineering and the M.S. and Ph.D. degrees in electrical engineering and computer science from Seoul National University, Seoul, South Korea, in 1999, 2001, and 2005, respectively. From 2005 to 2009, he was a Research Staff Member with the Samsung Advanced Institute of Technology, Samsung Electronics Company Ltd. In 2009, he joined the School of Electrical and Computer Engineering, Ulsan National Institute of Science and Technology, Ulsan, South Korea, where he is currently an Associate Professor. His research interests include image, video, and 3-D visual processing, computer vision, and multimedia data compression.