ROBUST CAMERA POSE ESTIMATION FOR IMAGE STITCHING

Laixi Shi[1,2]*, Dehong Liu*[2]*, Jay Thornton*[2]

[1] Carnegie Mellon University, PA, USA
[2] Mitsubishi Electric Research Laboratories (MERL), Cambridge, MA, USA

2021 IEEE International Conference on Image Processing (ICIP) | 978-1-6654-4115-5/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICIP42928.2021.9506680

ABSTRACT

Camera pose estimation plays a crucial role in stitching overlapped images captured by a camera to achieve a broad view of interest. In this paper, we propose a robust camera pose estimation approach to stitching images of a large 3D surface with known geometry. In particular, given a collection of images, we first construct a relative pose matrix estimation of all image pairs from the collection, where each entry of the matrix is calculated by solving a perspective-n-point (PnP) problem over the corresponding pair of images. To continue, we jointly estimate all camera poses by solving an optimization problem that exploits the underlying rank-2 relative pose matrices and the joint sparsity of camera pose errors. Finally, images are projected on to the 3D surface of interest based on estimated camera poses for the subsequent stitching process. Numerical experiments demonstrate that our proposed method outperforms existing methods in terms of both camera pose estimation and image stitching quality.

Index Terms— Camera pose estimation, low rank matrix, joint sparsity, image stitching

1. INTRODUCTION

Image stitching techniques aim to fuse a collection of overlapped images to achieve a broad view of an object or a scene of interest with satisfactory resolution in applications such as Google earth mapping [1], video stabilization [2], panoramic image construction [3], and 3D structure reconstruction [4], *etc*. In order to achieve the high stitching precision required by numerous applications, it is necessary to calibrate geometric distortions of collected images, and for this camera pose estimation for each image is crucial, as tiny errors in camera pose estimation may lead to evident distortion and visual mismatch of the stitched image, and consequently restrict subsequent applications. Therefore, it is of great importance and a daunting challenge to obtain sufficiently accurate estimates of camera poses for a given collection of images.

In recent decades, pairwise image alignment and stitching methods have been widely explored by using both feature-based and pixel-based methods [2, 5, 6]. In particular, for feature-based methods, images can be stitched one by one by aligning a new image to the previous image using SIFT-feature [7] matching points of this image pair. Pairwise image stitching methods generally work well for a small number of images with explicit matching points. However, it may fail when the total number of images is large, and feature points are not well matched, leaving outliers that cause an abnormal camera pose error that will l propagate to subsequent camera pose estimates. The accumulated pose error deteriorates the image stitching performance. To solve this problem, global alignment approaches such as bundle adjustment have been explored to promote globally consistency between all the images and improve the robustness of stitching [2]. To estimate the camera pose for each target image to be stitched, bundle adjustment considers feature matching points between the target image and a bundle of its overlapping reference images instead of just one reference image. Therefore, bundle adjustment shows robustness by comprehensively exploiting the matching point with all neighbors. However, the performance may still not be satisfactory, especially when with high precision stitching is required, since even a few matching point outliers can cause irreversible camera pose errors that accumulate during the stitching process.

In this paper, we propose a robust camera pose estimation method which accurately estimates camera poses even with abnormal pairwise camera pose estimation errors, when a large collection of images are being fused. Given a sequence of indexed images, we first estimate the pairwise relative pose matrix, in which each entry corresponds to a target and a reference image pair indexed by the row and the column, respectively. For each overlapping image pair, the corresponding entry is achieved by solving a perspective-n-point (PnP) problem using feature matching points in the overlapped area, otherwise the entry is non-observable. As a global alignment approach, our proposed method aims to jointly estimate all camera poses based on the partially observed relative pose matrix estimates, exploiting the underlying rank-2 relative pose matrices and the joint sparsity of the pose estimation errors [8].

Low-rankness and sparsity have been widely studied in recent years to exploit the low-dimensional structure of signals, such as for image restoration [9], signal processing [10], video processing [11–14], machine learning [15, 16], and Euclidean distance matrix reconstruction [17]. A well-known method promoting those properties is robust principal component analysis (RPCA) [18], which recovers a low-rank matrix and a sparse error matrix from their superpositioned observations. Inspired by RPCA, here we exploit the low-rankness in an implicit way rather than using an explicit low-rank constraint such as matrix nuclear norm. We also promote the joint sparsity of relative pose errors and enable the estimation in the case of partially observed matrices. Our main contributions are summarized as:

1. We proposed a robust image stitching method with a formulation promoting the underlying rank-2 matrices of relative poses and the joint sparsity of camera pose errors, which exhibits robustness even under abnormal camera poses.

2. We explored the general matrix reconstruction problem given partial observations, when the matrix has specific underlying structure. The proposed method is potentially applicable to various other scenarios targeting a sequence of linearly related estimations.

3. We conducted numerical experiments and demonstrated that our method significantly outperforms existing methods.

This work was finished when Laixi Shi was an intern at MERL.

2. ROBUST IMAGE GLOBAL STITCHING

2.1. Image Acquisition Model

In this work, we target the problem of stitching images of a large 3D surface with known geometry. Without loss of generality, we consider a collection of N images $\{\boldsymbol{X}_n\}_{n=1}^N$ covering a large surface $\boldsymbol{U}$, with each image overlapping with its neighbors. For instance, we consider a huge painting surface $\boldsymbol{U}$, as illustrated in Fig.1.

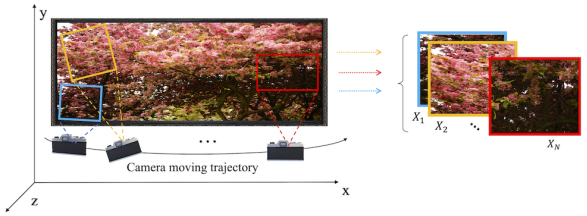


Fig. 1. Schematic diagram of image acquisition.

We consider a six-degrees-of-freedom (6-DoF) pin-hole camera model in image acquisition [2]. Based on this model, each pixel (x_c, y_c) in image $\boldsymbol{X}_n$ is projected from a point (x_u, y_u, z_u) on the 3D surface $\boldsymbol{U}$ according to

$$\begin{bmatrix} x_c \\ y_c \\ 1 \end{bmatrix} = \frac{1}{v} \begin{bmatrix} f & 0 & 0 & 0 \\ 0 & f & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{R} & \boldsymbol{T} \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x_u \\ y_u \\ z_u \\ 1 \end{bmatrix}, \tag{1}$$

where $\boldsymbol{R} = \boldsymbol{R}_x(\theta_x)\boldsymbol{R}_y(\theta_y)\boldsymbol{R}_z(\theta_z) \in \mathbb{R}^{3\times 3}$ is the rotation matrix with three DoFs determined by Euler angles θ_x, θ_y, and θ_z, with $\boldsymbol{R}_x(\cdot)$, $\boldsymbol{R}_y(\cdot)$, and $\boldsymbol{R}_z(\cdot)$ representing the rotation matrix around x, y, and z axes, respectively [19]; $\boldsymbol{T} = [T_x, T_y, T_z]^\top \in \mathbb{R}^3$ is the translation vector of the translations in x, y, and z directions; f is the focal length of the camera; and v is a pixel-dependent scaling factor to ensure that the projected pixel lies in the focal plane. We note that the pose of the 6-DoF camera is determined by a vector $\boldsymbol{p} = [\theta_x, \theta_y, \theta_z, T_x, T_y, T_z]^\top \in \mathbb{R}^K$ with $K = 6$.

2.2. Relative Camera Pose Matrices

For image stitching, an essential task is to estimate camera poses $\{\boldsymbol{p}_n\}_{n=1}^N$ for all images $\{\boldsymbol{X}_n\}_{n=1}^N$ in the collection such that the projective transformation of surface points shown in (1) can be properly calibrated. To simplify notations, we concatenate camera pose vectors $\{\boldsymbol{p}_n\}_{n=1}^N$ as row vectors into a matrix $\boldsymbol{P}$ and denote the k-th column (k-th dimension of all poses) vector of $\boldsymbol{P}$ as $\boldsymbol{h}_k$, *i.e.*,

$$\boldsymbol{P} = [\boldsymbol{p}_1, \boldsymbol{p}_2, \cdots, \boldsymbol{p}_N]^\top = [\boldsymbol{h}_1, \boldsymbol{h}_2, \cdots, \boldsymbol{h}_K] \in \mathbb{R}^{N\times K}. \tag{2}$$

For each k-th parameter of the camera poses, we define a relative camera pose matrix $\boldsymbol{L}^{(k)} \in \mathbb{R}^{N\times N}$ with each entry

$$\begin{aligned} \boldsymbol{L}^{(k)}(i,j) =& \boldsymbol{h}_k(i) - \boldsymbol{h}_k(j), \\ & \text{for} \quad i,j = 1,2,\cdots,N; k = 1,2,\cdots,K, \end{aligned} \tag{3}$$

where $\boldsymbol{h}_k(i)$ denotes the i-th parameter of $\boldsymbol{h}_k$, and $\boldsymbol{L}^{(k)}(i,j)$, the i-th row and j-th column entry of $\boldsymbol{L}^{(k)}$, represents the k-th parameter of the relative camera pose associated with the image pair $\boldsymbol{X}_i$ and $\boldsymbol{X}_j$. A key observation is that the matrix $\boldsymbol{L}^{(k)}$ can be rewritten as

$$\boldsymbol{L}^{(k)} = \boldsymbol{h}_k \mathbf{1}^\top - \mathbf{1}\boldsymbol{h}_k^\top \in \mathbb{R}^{N\times N}, \tag{4}$$

where $\mathbf{1}$ is an N-dimensional vector with all entries equal to 1. It is straightforward to verify that for any real vector $\boldsymbol{h}_k \neq a\mathbf{1}$, where a is a real scalar, we have

$$\text{rank}(\boldsymbol{L}^{(k)}) = \text{rank}(\boldsymbol{h}_k\mathbf{1}^\top) + \text{rank}(\mathbf{1}\boldsymbol{h}_k^\top) = 2. \tag{5}$$

2.3. Camera Pose Estimation

Motivated by the aforementioned rank-2 property of the relative camera pose matrices, we aim to recover camera poses from the relative camera pose matrices by exploiting the low-rank structure. To this end, we first construct an estimate of each relative camera pose matrix $\widetilde{\boldsymbol{L}}^{(k)} \in \mathbb{R}^{N\times N}$.

For each entry $\widetilde{\boldsymbol{L}}^{(k)}(i,j)$, we calculate the relative camera pose between the i-th image and the j-th image (reference) using a pairwise image stitching method. With the 3D surface geometry prior and the pose of the reference image, it can be modeled as a perspective-n-point (PnP) problem using SIFT matching feature points [5, 20–24]. For those image pairs that are not overlapped, due to unavailability of matching feature points, their corresponding entries in $\widetilde{\boldsymbol{L}}^{(k)}$ are missing. Therefore, we introduce a binary matrix $\boldsymbol{M} \in \mathbb{R}^{N\times N}$ indicating the observability of entries in $\widetilde{\boldsymbol{L}}^{(k)}$, where $\boldsymbol{M}(i,j) = 1$ if $\widetilde{\boldsymbol{L}}^{(k)}(i,j)$ is observable, otherwise $\boldsymbol{M}(i,j) = 0$. Considering that matching point outliers may lead to abnormal pose estimation errors, we model the observed noisy relative pose matrix as

$$\widetilde{\boldsymbol{L}}^{(k)} = \boldsymbol{L}^{(k)} + \boldsymbol{S}^{(k)}, \quad \text{for } k = 1,\cdots,K, \tag{6}$$

where $\boldsymbol{S}^{(k)} \in \mathbb{R}^{N\times N}$ represents sparse pose estimation errors. To introduce our proposed approach, we first vectorize all *observed* entries (according to $\boldsymbol{M}$) of $\widetilde{\boldsymbol{L}}^{(k)}$, $\boldsymbol{S}^{(k)}$ and $\boldsymbol{L}^{(k)}$ respectively as

$$\begin{aligned} \widetilde{\boldsymbol{l}}^{(k)} &= \text{vec}(\{\widetilde{\boldsymbol{L}}^{(k)}(i,j)|_{\boldsymbol{M}(i,j)=1}\}) \in \mathbb{R}^{|\boldsymbol{M}|}, \\ \boldsymbol{s}^{(k)} &= \text{vec}(\{\boldsymbol{S}^{(k)}(i,j)|_{\boldsymbol{M}(i,j)=1}\}) \in \mathbb{R}^{|\boldsymbol{M}|}, \\ \boldsymbol{l}^{(k)} &= \text{vec}(\{\boldsymbol{L}^{(k)}(i,j)|_{\boldsymbol{M}(i,j)=1}\}) \in \mathbb{R}^{|\boldsymbol{M}|}. \end{aligned} \tag{7}$$

where $\widetilde{\boldsymbol{l}}^{(k)}, \boldsymbol{s}^{(k)}, \boldsymbol{l}^{(k)}$ are with the same element order, and $|\boldsymbol{M}|$ is the cardinality of the nonzero entries in $\boldsymbol{M}$.

Recalling (4), we can express $\boldsymbol{l}^{(k)}$ as

$$\boldsymbol{l}^{(k)} = \boldsymbol{A}\boldsymbol{h}_k, \tag{8}$$

with

$$\boldsymbol{A} = [\cdots, \boldsymbol{\alpha}_{i,j}, \cdots]^\top \in \mathbb{R}^{|\boldsymbol{M}|\times N}, \tag{9}$$

where $(i,j) \in \{(i,j) : \boldsymbol{M}(i,j) = 1\}$. Vector $\boldsymbol{\alpha}_{i,j} \in \mathbb{R}^N$ has all-zero entries except that $\boldsymbol{\alpha}_{i,j}(i) = 1$ and $\boldsymbol{\alpha}_{i,j}(j) = -1$. Concatenating those of all the K parameters, we denote

$$\begin{aligned} \widetilde{\boldsymbol{L}} &= [\widetilde{\boldsymbol{l}}^{(1)}, \widetilde{\boldsymbol{l}}^{(2)}, \cdots, \widetilde{\boldsymbol{l}}^{(K)}] \in \mathbb{R}^{|\boldsymbol{M}|\times K}, \\ \boldsymbol{S} &= [\boldsymbol{s}^{(1)}, \boldsymbol{s}^{(2)}, \cdots, \boldsymbol{s}^{(K)}] \in \mathbb{R}^{|\boldsymbol{M}|\times K}, \\ \boldsymbol{L} &= [\boldsymbol{l}^{(1)}, \boldsymbol{l}^{(2)}, \cdots, \boldsymbol{l}^{(K)}] = \boldsymbol{A}\boldsymbol{P} \in \mathbb{R}^{|\boldsymbol{M}|\times K}. \end{aligned}$$

Considering that different parameters of the camera pose (such as rotation angle θ_x and transition T_x) are of different magnitude ranges, we introduce a vector $\boldsymbol{w} = [w_1, w_2, \cdots, w_K]^\top$, where $w_k = \frac{N-1}{\sum_{i=1}^{N-1}|\widetilde{\boldsymbol{L}}^{(k)}(i,i+1)|}$, to normalize the magnitude ranges. Consequently, the camera pose estimation problem is formulated as

$$\min_{\boldsymbol{S},\boldsymbol{P}} \frac{1}{2} \left\| \left(\widetilde{\boldsymbol{L}} - \boldsymbol{A}\boldsymbol{P} - \boldsymbol{S} \right) \boldsymbol{W} \right\|_F^2 + \lambda \left\| \boldsymbol{S}\boldsymbol{W} \right\|_{2,1}, \tag{10}$$

where $\boldsymbol{W} = \text{diag}\{\boldsymbol{w}\}$ is a diagonal matrix whose diagonal is $\boldsymbol{w}$, $\|\cdot\|_F$ denotes the Frobenius norm, $\|\cdot\|_{2,1}$ denotes the $\ell_{2,1}$ norm of a matrix defined as $\|\boldsymbol{Q}\|_{2,1} = \sum_{i=1}^{I} \sqrt{\sum_{j=1}^{J} [Q(i,j)]^2}$ for any matrix $\boldsymbol{Q} \in \mathbb{R}^{I \times J}$ [8], and λ is the regularization parameter.

Note that (10) is similar to the well-known RPCA [18] which decomposes the observation matrix into a low-rank matrix and a sparse noise matrix. Nevertheless, here we implicitly impose low-rankness rather than use an explicit low-rank constraint such as nuclear-norm minimization. In addition, we promote joint sparsity across the noise of different parameters of each pose using the $\ell_{2,1}$ norm. It is inspired by the observation that when one parameter of a camera pose is not well estimated, all other parameters exhibit large errors with high probability.

To solve (10), we utilize the alternating minimization method by alternately updating $\boldsymbol{S}$ and $\boldsymbol{P}$ until its stopping criteria is satisfied. The sub-problem with respect to $\boldsymbol{P}$ is a standard least-squares problem, which can be solved by

$$\boldsymbol{P}^{(t)} = \boldsymbol{A}^{\dagger} \left(\widetilde{\boldsymbol{L}} - \boldsymbol{S}^{(t-1)} \right), \tag{11}$$

where $\boldsymbol{A}^{\dagger}$ denotes the pseudo-inverse of $\boldsymbol{A}$, and superscripts (t) and $(t-1)$ represent the number of iterations during the process.

The subproblem of $\boldsymbol{S}$ can be solved by row-dependent soft-thresholding as

$$\boldsymbol{S}_{i,:}^{(t)} = (\widetilde{\boldsymbol{L}} - \boldsymbol{A}\boldsymbol{P}^{(t)})_{i,:} \odot \max\left(0, 1 - \frac{\lambda}{\left\| (\widetilde{\boldsymbol{L}} - \boldsymbol{A}\boldsymbol{P}^{(t)})_{i,:} \boldsymbol{W} \right\|_2}\right), \tag{12}$$

where the subscript $(i,:)$ denotes the i-th row of the corresponding matrix, $\odot$ stands for the element-wise product. The overall algorithm is summarized in Alg. 1.

Algorithm 1: Robust Camera Pose Estimation

Input: Constructed relative pose matrix estimation $\widetilde{\boldsymbol{L}}$, the mask $\boldsymbol{M}$, and $\boldsymbol{A}$.
Initial $\boldsymbol{S}^{(0)} = \boldsymbol{0}$, $\boldsymbol{P}^{(0)} = \boldsymbol{0}$, parameters λ, ϵ, and $iter$;
for $t \leftarrow 1$ **to** $iter$ **do**
  Update $\boldsymbol{P}^{(t)} \leftarrow$ *solve* (11) *with* $\boldsymbol{S}^{(t-1)}$*;*
  Update $\boldsymbol{S}^{(t)} \leftarrow$ *solve* (12) *with* $\boldsymbol{P}^{(t)}$*;*
  if $\left\| \boldsymbol{P}^{(t)} - \boldsymbol{P}^{(t-1)} \right\|_F \leq \epsilon$ **then** break;
end
Output: $\hat{\boldsymbol{P}} \leftarrow \boldsymbol{P}^{(t)}$.

It is clear that when $\lambda = 0$ and $\boldsymbol{S} = \boldsymbol{0}$, Alg. 1 is reduced to the least-squares method for camera pose estimation

$$\widehat{\boldsymbol{P}}_{LS} = \boldsymbol{A}^{\dagger} \widetilde{\boldsymbol{L}}. \tag{13}$$

2.4. Image stitching

Note that any global rigid translation on camera poses leads to the same relative camera pose matrix $\boldsymbol{L}$ [25]. To avoid ambiguity, we take the first camera pose as a noiseless reference, with other camera poses calculated from the output $\widehat{\boldsymbol{P}}$ of Alg. 1 as follows

$$\widehat{\boldsymbol{p}}_n = \widehat{\boldsymbol{P}}_{n,:}^{\top} - \widehat{\boldsymbol{P}}_{1,:}^{\top} + \boldsymbol{p}_1, \ \text{ for } \ n = 1, \ldots N. \tag{14}$$

Once we obtain estimated camera poses $\{\widehat{\boldsymbol{p}}_n\}_{n=1}^{N}$, we project all N images to the 3D surface $\boldsymbol{U}$ according to (1) to construct the final stitched image using interpolation.

3. NUMERICAL EXPERIMENTS

3.1. Experimental Settings

To examine our proposed joint sparsity and rank-2-based global stitching approach, we simulated a sequence of $N = 50$ images $\{\boldsymbol{X}_n\}_{n=1}^{N}$ of size 500×600 captured by a camera at different poses using model shown in (1), as illustrated in Fig. 1. To ensure a full coverage of the huge painting surface $\boldsymbol{U}$, the camera moves in a linear-scan pattern, capturing overlapped images, with each image covering a small portion of the whole surface. Because of the random perturbation of poses during image collection, camera poses must be estimated in the stitching process.

To estimate camera poses, we compare four different methods: (1) pairwise method, (2) bundle adjustment method, (3) least-squares method, and (4) our proposed method.

For the pairwise method, we utilize a SIFT-feature-based perspective-n-point (PnP) method [26, 27] to sequentially estimate all camera poses with prior knowledge of the 3D scene geometry. For the bundle adjustment baseline, we expand SIFT-feature matching points by including all those between the target image and its overlapping neighbors. Specifically, when stitching the i-th image, we consider the set of neighbors $\mathcal{G}_i = \{g | g < i, |g - i| \leq 25\}$ for feature matching points construction[1]. For the least-squares method and our proposed method, the mask $\boldsymbol{M}$ is given according to the valid number of overlapping neighbors by

$$\boldsymbol{M}(i,j) = \begin{cases} 1 & \text{if } |i-j| \leq 25, i \neq j \\ 0 & \text{otherwise .} \end{cases} \tag{15}$$

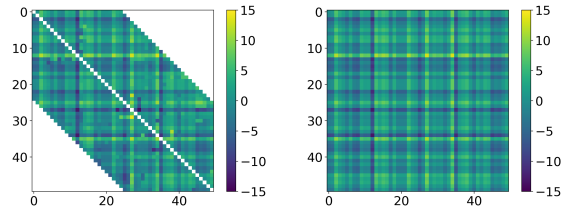
The hyper-parameter in (10) is tuned as $\lambda = 0.01$ for this specific application.

3.2. Experimental Results

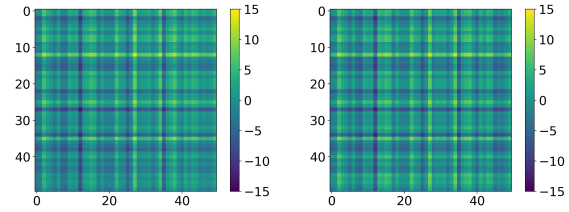
To illustrate our observation with missing data, in Fig. 2 (a) we illustrate one partially observed relative pose matrix $\widetilde{\boldsymbol{L}}^{(4)} \in \mathbb{R}^{N \times N}$ as an example, where the white area in the square matrix corresponds to the missing values. For comparison, in Fig. 2 (b) we show the underlying ground truth matrix $\boldsymbol{L}^{(4)}$ that we desire to recover. Since $\boldsymbol{L}^{(4)}$ is a rank-2 matrix, we can clearly observe similar patterns in rows, and also in columns. Furthermore, Fig. 2 (c) and (d) are the estimated relative pose matrices $\widehat{\boldsymbol{L}}^{(4)}$ via the least-squares baseline and our approach respectively. Although both estimates exhibit patterns of low-rankness, the result using our proposed method preserves the true pattern with high fidelity regardless of the interference of large relative pose observation errors in Fig. 2 (a).

To verify the joint sparsity of pose errors using the pairwise stitching method, we present two ground truth noise matrices $\boldsymbol{S}^{(4)}$ and $\boldsymbol{S}^{(3)}$ in Fig.3 (a) and (b) respectively. The two plots imply that the pattern of the sparse errors with respect to different pose parameters are consistent, which confirms our model assumption. Due to space limit, we only show one estimated sparse error matrix $\widehat{\boldsymbol{S}}^{(4)}$ in Fig.3(c) and note it accurately recovered the ground truth

[1] Although we could also use neighbors $\mathcal{G}_i' = \{g | \, |g - i| \leq 25\}$ for the bundle method, we utilize $\mathcal{G}_i$ instead of $\mathcal{G}_i'$ because $\mathcal{G}_i$ gives the bundle method better results.



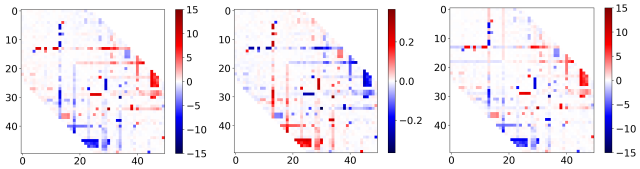
(a) Partially observed $\widetilde{\boldsymbol{L}}^{(4)}$ (b) Ground truth $\boldsymbol{L}^{(4)}$



(c) $\widehat{\boldsymbol{L}}^{(4)}$ via **LS** (d) $\widehat{\boldsymbol{L}}^{(4)}$ via Ours

Fig. 2. Examples of relative camera pose matrices. (a) Partially observed matrix using PnP method, (b) Underlying rank-2 ground truth, (c) Recovered matrix via the least-squares method, and (d) Recovered matrix via our proposed approach.

$\boldsymbol{S}^{(4)}$ shown in Fig.3(a), which illustrates that our proposed approach enables abnormal pose error detection.



(a) $\boldsymbol{S}^{(4)}$ (b) $\boldsymbol{S}^{(3)}$ (c) $\widehat{\boldsymbol{S}}^{(4)}$ via Ours

Fig. 3. Example sparse error matrices of relative camera pose matrices. (a) Sparse error matrix $\boldsymbol{S}^{(4)} = (\widetilde{\boldsymbol{L}}^{(4)} - \boldsymbol{L}^{(4)}) \odot \boldsymbol{M}$, (b) Sparse error matrix $\boldsymbol{S}^{(3)} = (\widetilde{\boldsymbol{L}}^{(3)} - \boldsymbol{L}^{(3)}) \odot \boldsymbol{M}$, and (c) Recovered sparse pose error matrix $\widehat{\boldsymbol{S}}^{(4)}$ using our proposed method.

To quantitatively analyze the performance of camera pose estimation, we compute the (average) relative estimation error $e = \frac{1}{K}\sum_{k=1}^{K} \frac{\|\boldsymbol{t}^{(k)} - \widehat{\boldsymbol{t}}^{(k)}\|_2}{\|\boldsymbol{t}^{(k)}\|_2}$ of different methods, where $\widehat{\boldsymbol{t}}^{(k)}$ is the reconstruction of $\boldsymbol{t}^{(k)}$. We display the relative errors with respect to the image number N in Fig. 4, which shows that under random abnormal camera pose estimation errors, the camera pose errors of both pairwise and bundle methods gradually accumulate across images. In comparison, although the least-squares method maintains the relative error to a lower level, our proposed method reduces the error even more. The robustness of our method to abnormal pose estimation error is achieved by promoting the joint sparsity of the noise matrix.

Using the estimated camera poses, we project images on the 3D surface and interpolate the projected pixels to get the final image of the large surface $\widehat{\boldsymbol{U}}$. Fig. 5 shows (a) the ground truth image, (b) the stitched image using the least-squares method (best baseline), and (c) the image using our proposed method for comparison. Fig. 5 demonstrates that both the least-squares method and our proposed method achieved visually satisfying quality compared to the ground truth. However, when we zoom into a small area, a lot of artifacts are observed in the stitched image using the least-squares method, as shown in Fig.5(e) compared to the ground truth shown in Fig. 5(d). Our proposed method significantly reduces these stitching artifacts as shown in Fig.5(f).

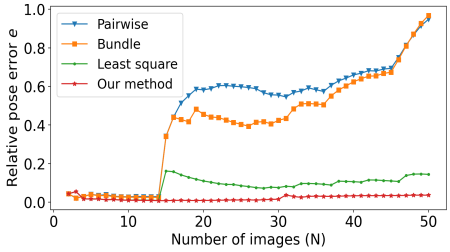
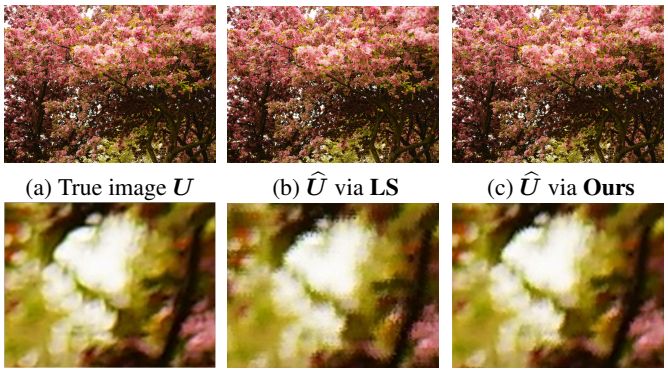


Fig. 4. The relative error of reconstructed relative pose matrix as the number of images N increases.



(a) True image $\boldsymbol{U}$ (b) $\widehat{\boldsymbol{U}}$ via **LS** (c) $\widehat{\boldsymbol{U}}$ via **Ours**

(d) Local area of (a) (e) Local area of (b) (f) Local area of (c)

Fig. 5. Comparison of images between (a) ground truth $\boldsymbol{U}$, (b) stitched image via the least-squares method, and (c) stitched image via our proposed approach. Images (d), (e), and (f) are the same local area of (a), (b), and (c), respectively.

To further evaluate the performance of different stitching methods, we compute the relative error of relative camera poses and the Peak-Signal-to-Noise Ratios (PSNR) of their stitched images $\widehat{\boldsymbol{U}}$, as shown in Table. 1. It is clear that our proposed method outperforms other baselines, achieving a much smaller relative camera pose estimation error and a much larger PSNR of the stitched image, by exploiting joint sparsity and low-rankness properties.

	Pairwise	Bundle	LS	Ours
Relative pose error	0.948	0.968	0.140	**0.037**
PSNR of $\widehat{\boldsymbol{U}}$ (dB)	19.23	20.94	26.68	**30.29**

Table 1. Relative errors of the relative camera pose matrices and PSNRs of stitching results $\widehat{\boldsymbol{U}}$ using $N = 50$ images.

4. CONCLUSION

We propose a robust camera pose estimation method for stitching a large collection of images of a 3D surface with known geometry. To address the issue of accumulating camera pose error using existing methods, we constructed a partially observed relative pose matrix for each parameter of camera poses, and decompose it into a rank-2 matrix of relative camera poses and a sparse matrix of camera pose errors by exploiting the joint sparsity of camera pose errors in estimating camera poses. Numerical experiments with simulating images captured by a camera with random pose perturbations causing abnormal pair wise image pose estimation errors demonstrate that our proposed method is capable of yielding robust camera pose estimates and significantly better stitching results than popular baseline methods.

5. REFERENCES

[1] Irwin Scollar, "Google earth: improving mapping accuracy," *AARGnews*, vol. 20, no. 2, pp. 19, 2013.

[2] Richard Szeliski, "Image alignment and stitching: A tutorial," *Foundations and Trends® in Computer Graphics and Vision*, vol. 2, no. 1, pp. 1–104, 2006.

[3] Matthew Brown and David G Lowe, "Automatic panoramic image stitching using invariant features," *International journal of computer vision*, vol. 74, no. 1, pp. 59–73, 2007.

[4] Johannes L Schonberger and Jan-Michael Frahm, "Structure-from-motion revisited," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4104–4113.

[5] Roy Sheffer and Ami Wiesel, "PnP-Net: A hybrid perspective-n-point network," *arXiv preprint arXiv:2003.04626*, 2020.

[6] Lyu Wei, Zhou Zhong, Chen Lang, and Zhou Yi, "A survey on image and video stitching," *Virtual Reality & Intelligent Hardware*, vol. 1, no. 1, pp. 55–83, 2019.

[7] David G Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[8] Yi Yang, Heng Tao Shen, Zhigang Ma, Zi Huang, and Xiaofang Zhou, "$\ell_{2,1}$-norm regularized discriminative feature selection for unsupervised learning," in *IJCAI international joint conference on artificial intelligence*, 2011.

[9] Yao Hu, Debing Zhang, Jieping Ye, Xuelong Li, and Xiaofei He, "Fast and accurate matrix completion via truncated nuclear norm regularization," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 35, no. 9, pp. 2117–2130, 2012.

[10] Guangcan Liu, Zhouchen Lin, Shuicheng Yan, Ju Sun, Yong Yu, and Yi Ma, "Robust recovery of subspace structures by low-rank representation," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 171–184, 2012.

[11] Thierry Bouwmans, Sajid Javed, Hongyang Zhang, Zhouchen Lin, and Ricardo Otazo, "On the applications of robust PCA in image and video processing," *Proceedings of the IEEE*, vol. 106, no. 8, pp. 1427–1457, 2018.

[12] Namrata Vaswani, Thierry Bouwmans, Sajid Javed, and Praneeth Narayanamurthy, "Robust subspace learning: Robust pca, robust subspace tracking, and robust subspace recovery," *IEEE signal processing magazine*, vol. 35, no. 4, pp. 32–55, 2018.

[13] Thierry Bouwmans, Namrata Vaswani, Paul Rodriguez, René Vidal, and Zhouchen Lin, "Introduction to the issue on robust subspace learning and tracking: theory, algorithms, and applications," *IEEE Journal of Selected Topics in Signal Processing*, vol. 12, no. 6, pp. 1127–1130, 2018.

[14] Namrata Vaswani, Yuejie Chi, and Thierry Bouwmans, "Rethinking pca for modern data sets: Theory, algorithms, and applications [scanning the issue]," *Proceedings of the IEEE*, vol. 106, no. 8, pp. 1274–1276, 2018.

[15] Rongkai Xia, Yan Pan, Lei Du, and Jian Yin, "Robust multi-view spectral clustering via low-rank and sparse decomposition," in *Proceedings of the twenty-eighth AAAI conference on artificial intelligence*, 2014, pp. 2149–2155.

[16] Joonseok Lee, Seungyeon Kim, Guy Lebanon, and Yoram Singer, "Local low-rank matrix approximation," in *International conference on machine learning*, 2013, pp. 82–90.

[17] Dehong Liu, Hassan Mansour, Petros T Boufounos, and Ulugbek S Kamilov, "Robust sensor localization based on euclidean distance matrix," in *2018 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 2018, pp. 7998–8001.

[18] Emmanuel J Candès, Xiaodong Li, Yi Ma, and John Wright, "Robust principal component analysis?," *Journal of the ACM (JACM)*, vol. 58, no. 3, pp. 1–37, 2011.

[19] Mike Day and Insomniac Games, "Extracting euler angles from a rotation matrix," *Insomniac Games R&D. Available online at: http://www. insomniacgames. com/mike-day-extracting-euler-angles-from-a-rotation-matrix*, 2012.

[20] Vincent Lepetit, Francesc Moreno-Noguer, and Pascal Fua, "EPnP: An accurate O(n) solution to the PnP problem," *International journal of computer vision*, vol. 81, no. 2, pp. 155, 2009.

[21] C-P Lu, Gregory D Hager, and Eric Mjolsness, "Fast and globally convergent pose estimation from video images," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 22, no. 6, pp. 610–622, 2000.

[22] Yinqiang Zheng, Yubin Kuang, Shigeki Sugimoto, Kalle Astrom, and Masatoshi Okutomi, "Revisiting the PnP problem: A fast, general and optimal solution," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 2344–2351.

[23] Luis Ferraz, Xavier Binefa, and Francesc Moreno-Noguer, "Very fast solution to the PnP problem with algebraic outlier rejection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 501–508.

[24] Yinqiang Zheng, Shigeki Sugimoto, and Masatoshi Okutomi, "ASPnP: An accurate and scalable solution to the perspective-n-point problem," *IEICE Trans. on Information and Systems*, vol. 96, no. 7, pp. 1525–1535, 2013.

[25] Ivan Dokmanic, Reza Parhizkar, Juri Ranieri, and Martin Vetterli, "Euclidean distance matrices: essential theory, algorithms, and applications," *IEEE Signal Processing Magazine*, vol. 32, no. 6, pp. 12–30, 2015.

[26] Yanting Ma, Dehong Liu, Hassan Mansour, Ulugbek S Kamilov, Yuichi Taguchi, Petros T Boufounos, and Anthony Vetro, "Fusion of multi-angular aerial images based on epipolar geometry and matrix completion," in *2017 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2017, pp. 1197–1201.

[27] Laixi Shi, Dehong Liu, Masaki Umeda, and Norihiko Hana, "Fusion-based digital image correlation framework for strain measurement," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 1400–1404.