2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
September 27 - October 1, 2021. Prague, Czech Republic

Direct Bundle Adjustment for 3D Image Fusion with Application to Transesophageal Echocardiography*

Zhehua Mao[1], Liang Zhao[1(✉)], Shoudong Huang[1], Yiting Fan[2], Alex Pui-Wai Lee[3]

***Abstract*— In this paper, we propose a novel algorithm for fusing a sequence of 3D images, named as Direct Bundle Adjustment (DBA). This algorithm simultaneously optimizes the global pose parameters of image frames and the intensity values of the fused global image using the 3D image data directly (without extracting features from the images). This one-step 3D image fusion approach is achieved by formulating the problem as an optimization problem to minimize the intensity differences between the global image and the corresponding points in the different local images. The proposed DBA method is particularly useful in the scenarios where distinct features are not available, such as Transesophageal Echocardiography (TEE) images. We validate the proposed method via simulated and in-vivo 3D TEE images. It is shown that the proposed method is robust to intensity noises and much more accurate than the conventional sequential fusion method.**

***Index Terms*— direct method, bundle adjustment, image fusion, 3D TEE, ultrasound imaging.**

I. Introduction

3D image registration and fusion is an important research topic that has attracted a lot of attention in the past decades [1], [2]. One important application of 3D image fusion is in the area of medical imaging, where a larger 3D image of a human organ can be obtained by fusing multiple smaller 3D medical images [3], [4]. As an example, 3D transesophageal echocardiography (TEE) image is a powerful tool for clinical imaging of the heart and is becoming increasingly popular in preoperative diagnosis and intraoperative guidance in the recent decade [5]. However, it suffers from a limited field of view (FoV) which makes it hard to visualize the entire heart without moving the transducer to different positions for image acquisition. Thus, the 3D fusion of TEE images has significant clinical value.

According to whether image registration/fusion requires the feature extraction from the images or not, image fusion can generally be divided into two categories: feature-based methods and direct (or intensity-based) methods [2]. Feature-based methods estimate the poses based on the extracted features and discard all the other information of the images. Thus, the accuracy of the pose estimate varies. When the environment contains abundant distinct features, feature-based methods can generate accurate pose estimation. However, when the number of distinct features in the environments is limited or when the feature extraction is difficult, the pose estimate can be very inaccurate. On the contrary, direct methods estimate the relative pose through maximizing the similarity between two images directly based on intensities of images, in which all the information from images can be used. The methods have occupied a dominant role in the field of medical registration [2], [6]. Commonly used similarity measures include Sum-of-Squared Differences (SSD), Normalized Cross-Correlation (NCC), and Mutual Information (MI) [7].

Most of the direct methods are designed for the registration and fusion of two images. The process involves designating one image as the reference (or fixed) image and registering the other image (moving image) to the reference one. After that, the moving image is fused with the reference image. When more than two images need to be fused, one common strategy is to use a sequential fusion method [8], [9]. This strategy, although being intuitive, has the issue of information reuse and will inevitably bring in accumulating errors in both the estimated pose parameters and the fused global image. Existing 3D TEE images fusion methods are all deduced from a sequence of pairwise registrations, thus the accuracy of the obtained global image (fused 3D image) is compromised [10]–[12]. Compared with pairwise methods, a better strategy is to optimize the poses of all local frames at the same time and distribute the errors evenly. [13] proposed to use multivariate similarity measures for registering a group of images. But this method may have problem with information reuse since the overlapping areas are taken into account in the objective function for multiple times, which also increase the extra complexity in the optimization. [14] proposed a congealing framework which estimates poses of local frames by minimising the sum of entropies of pixel values at each pixel location. As [15] mentioned, the use of entropy for registration in this method is problematic due to its poor optimization characteristics. And congealing involves high computational cost, thus sampling need to be used for 3D images [16].

Bundle adjustment (BA) [17] is regarded as Gold Standard for obtaining optimal frame poses of multiple 2D images in computer vision (CV) community. However, most of

*This work is supported by Australian Research Council (ARC) Discovery project "Visual Simultaneous Localisation and Mapping in Deformable Environments" (DP200100982).

[1]Zhehua Mao, Liang Zhao, and Shoudong Huang are with Centre for Autonomous Systems, Faculty of Engineering and Information Technology, University of Technology Sydney, Ultimo, NSW 2007, Australia (e-mails: Zhehua.Mao@student.uts.edu.au; Liang.Zhao@uts.edu.au; Shoudong.Huang@uts.edu.au).

[2]Yiting Fan is with Department of Cardiology, Shanghai Chest Hospital, Shanghai Jiao Tong University, Shanghai, China (e-mail: myricefyt@126.com).

[3]Alex Pui-Wai Lee is with Division of Cardiology, Department of Medicine and Therapeutics, Prince of Wales Hospital and Laboratory of Cardiac Imaging and 3D Printing, Li Ka Shing Institute of Health Science, Faculty of Medicine, The Chinese University of Hong Kong, Hong Kong, China (e-mail: alexpwlee@cuhk.edu.hk).

the BA algorithms are feature-based, which require feature correspondences between images prior to BA optimization. Although [18] and [19] proposed to used photometric information in their frameworks, their works require highly accurate initialization and are considered as refinement steps after feature-based methods. Importantly, one of their main objectives is still to optimize the positions of 3D scene points (or 3D shapes) besides camera poses, which are similar to many other volumetric/direct methods in CV community [20]–[22]. To our best knowledge, there is no BA method that can be used for multiple 3D image registration and fusion.

In this paper, we propose a direct method for simultaneously fusing multiple 3D images. In particular, we simultaneously optimize intensities of global image and the poses of local frames by minimizing the intensity differences between the global image and the corresponding points in different local frames. We call this method direct bundle adjustment (DBA). It should be noted that although the method is called DBA, it is quite different from most of the BA algorithms which belong to the feature-based methods, require feature correspondences between images, and are mostly for using 2D images to estimate the 3D feature positions (or shapes) in the reconstruction of the environments [17]–[19].

The novel DBA algorithm proposed in this paper has filled an important gap on 3D image registration and fusion. Without any reference image, correspondences, or information loss or reuse, the proposed DBA algorithm is an elegant way to obtain the optimal global image and poses of local frames. To our best knowledge, this is the first work of BA based on direct method for 3D image registration and fusion.

This paper is organized as follows. Section II provides the details of the problem formulation and methodology. To validate the proposed algorithm and assess its accuracy and robustness, in Section III, experiments are performed based on the simulated and in-vivo 3D TEE image data. It is shown that the proposed algorithm is robust to intensity noises and can obtain results much more accurate than the conventional sequential fusion method. Finally, Section IV concludes the paper and addresses some future work.

II. Methodology

A. Problem Statement

Fig. 1 illustrates a scenario in which there are m frames of 3D images and each image captures only a portion of the actual scene. We use the vector elements $\xi_i \in \mathbb{R}^{6\times 1}$ of Lie algebra [23] to represent the pose parameters of the probe at position i and the corresponding captured 3D image is represented as I_i. Suppose M is the global image which fuses all the information of local frames, it consists of n voxels and the intensity of one voxel p_j in M is denoted as $M(p_j)$ ($p_j = [u, v, w]^\top$, where u, v, w are integers which present a voxel location on the 3D grid). Assuming that a part of intensity information of voxel p_j in the global image is from a point p_{ij} in local frame I_i, in rigid scenario, transformation in Euclidean space from p_j to p_{ij} can be formally written as:

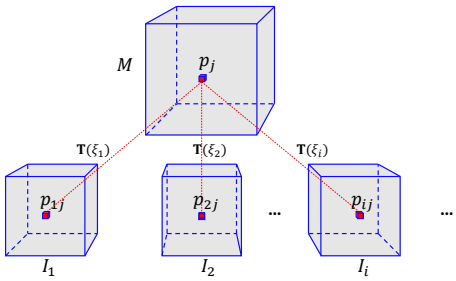

Fig. 1. Direct bundle adjustment (DBA): intensity of global image $\{M(p_j)\}$ and pose parameters of local frames $\{\xi_i\}$ are optimized simultaneously.

$$p_{ij} = f(\xi_i, p_j) = \mathbf{T}(\xi_i)p_j{}^{1}, \tag{1}$$

where $\mathbf{T}(\cdot) \in SE(3)$ maps the pose parameters ξ_i to a rigid transformation matrix in Euclidean space.

In this study, we seek to optimize the pose parameters of local frames and the intensities of voxels in the global image simultaneously. For a n-voxel global image M and m frames of 3D images, the overall state parameters considered in the proposed algorithm are:

$$\mathbf{X} = [\mathbf{X}_\xi^\top, \mathbf{X}_M^\top]^\top, \tag{2}$$

where $\mathbf{X}_\xi = [\xi_1^\top, ..., \xi_i^\top, ..., \xi_m^\top]^\top$ are pose parameters of the local frames, and $\mathbf{X}_M = [M(p_1), ..., M(p_j), ..., M(p_n)]^\top$ represent intensities of voxels in the global image M.

B. Formulation of Direct Bundle Adjustment

In Fig. 1, the intensity difference between $M(p_j)$ and $I_i(p_{ij})$ is:

$$\begin{aligned} e_{ij}(\xi_i, M(p_j)) &= M(p_j) - I_i(f(\xi_i, p_j)) \\ &= M(p_j) - I_i(p_{ij}), \end{aligned} \tag{3}$$

where p_j is constant and p_{ij} can be calculated from p_j using (1) during the optimization process. p_{ij} may not be integers thus not on the voxel of the grid of the local image. And the intensity of $I_i(p_{ij})$ is obtained using interpolation.

Based on the direct method, we propose the DBA method to obtain the optimal estimation of $\mathbf{X}$ by minimizing the sum of the squared intensity differences between global image and local images:

$$\hat{\mathbf{X}} = \underset{\mathbf{X}_\xi, \mathbf{X}_M}{\operatorname{argmin}} \sum_{j=1}^{n}\sum_{i=1}^{m} \sigma(p_{ij})(e_{ij}(\xi_i, M(p_j)))^2, \tag{4}$$

where $\sigma(p_{ij}) = 1$ if the transformed point p_{ij} located within the image region of I_i, otherwise $\sigma(p_{ij}) = 0$.

C. Solving the Optimization Problem

From (3), it is found that intensity difference e_{ij} is only dependent on ξ_i and intensity of voxel p_j in global image, i.e. $M(p_j)$. Therefore, according to the chain rule, we can

[1]We implicitly perform the conversion between 3D Euclidean coordinates and homogeneous coordinates for p_j and p_{ij} in the paper.

derive the partial derivatives of intensity difference e_{ij} w.r.t. ξ_i and $M(p_j)$:

$$\frac{\partial e_{ij}(\xi_i, M(p_j))}{\partial \xi_i} = -\frac{\partial I_i}{\partial f(\xi_i, p_j)} \frac{\partial f(\xi_i, p_j)}{\partial \xi_i}, \tag{5}$$

$$\frac{\partial e_{ij}(\xi_i, M(p_j))}{\partial (M(p_j))} = 1. \tag{6}$$

It is shown in (5) that the partial derivative of e_{ij} w.r.t. pose parameters ξ_i consists of two parts:

1) The first term $\partial I_i / \partial f$ in (5) is the partial derivative of intensity I_i w.r.t. f, which is a 1×3 vector.

2) The second term $\partial f / \partial \xi_i$ in (5) is the partial derivative of f w.r.t. pose parameters ξ_i, which is a 3×6 matrix.

Thus, the result of (5) will be a 1×6 vector. In addition, it is found in (6) that the partial derivative of intensity difference w.r.t. intensity at point p_j in the global image is a constant 1.

In our DBA algorithm, the state vector $\mathbf{X}$ to be optimized includes pose parameters of all local frames and intensities of all voxels in the global image. Thus, we can derive the Jacobian of one intensity difference e_{ij} w.r.t. state vector $\mathbf{X}$ as the following format:

$$J_{ij}(\mathbf{X}) = \left[0_{1\times 6}, ..., \frac{\partial e_{ij}}{\partial \xi_i}, ..., 0_{1\times 6}, 0, ..., \frac{\partial e_{ij}}{\partial (M(p_j))}, ...0 \right], \tag{7}$$

where $0_{1\times 6}$ denotes a 1×6 zero matrix.

If we write the overall intensity differences as a concatenation vector $e(\mathbf{X}) = [..., e_{ij}, ...]^\top$, the objective function $F(\mathbf{X})$ to be optimized can be defined as:

$$F(\mathbf{X}) = e(\mathbf{X})^\top e(\mathbf{X}). \tag{8}$$

According to (7), for all intensity differences $e(\mathbf{X})$, the overall Jacobian matrix can be built as a collection of J_{ij} as $J(\mathbf{X}) = [..., J_{ij}(\mathbf{X})^\top, ...]^\top$.

In this study, Gauss-Newton (GN) method is adopted to solve the nonlinear least-squares problem in (4). During GN iterations, step change $\Delta\mathbf{X}$ in each iteration can be calculated from the GN equation:

$$H \Delta \mathbf{X} = b, \tag{9}$$

where

$$H = J(\mathbf{X})^\top J(\mathbf{X}), \; b = -J(\mathbf{X})^\top e(\mathbf{X}). \tag{10}$$

D. Efficient Implementation

Sparsity and marginalization: GN method linearizes the nonlinear optimization problem (4). During the GN iteration, we need to solve the linear system (9). But this step is cumbersome because the size of matrix H in our problem is huge. For a case with m frames of 3D images and n voxels in the global image, the size of matrix H is $(6m+n)\times(6m+n)$, as shown in Fig. 2. And in matrix H, n is usually much larger than m. Take a 3D TEE image as an example, it usually contains millions of voxels in one single volume. Reconstructing a global image with several 3D TEE images, the size of matrix H is in the tens of millions multiply tens

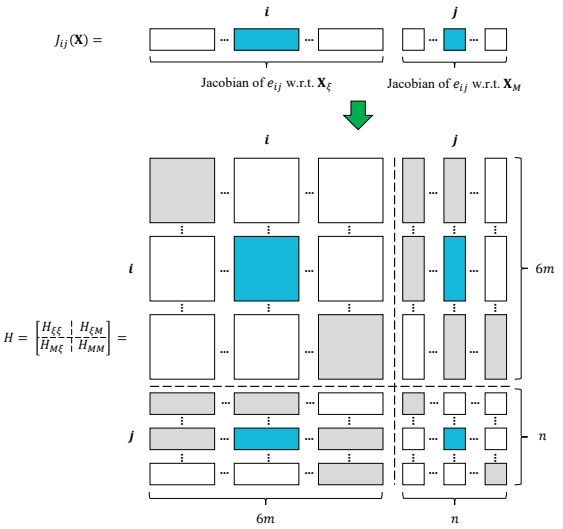

Fig. 2. Sparsity of matrix H: $H_{\xi\xi}$ is block diagonal and H_{MM} is diagonal. The size of H_{MM} is usually much larger than $H_{\xi\xi}$. The structure of $H_{\xi M}$ and $H_{M\xi}$ indicates the observation relationships between voxels in global image and local frames.

of millions. Calculating such a large matrix directly is time-consuming and impossible in most cases.

When inspecting (7), we find that $J_{ij}(\mathbf{X})$ is sparse and this sparse property will cause the sparsity in the matrix H finally. In Fig. 2, we use an intuitive structure to illustrate the sparse property of H. The upper sub-figure of $J_{ij}(\mathbf{X})$ is corresponding to the row vector in (7), the blocks with color in the figure means values in these blocks are nonzero. Since $J_{ij}(\mathbf{X})$ only has nonzero blocks at block i and j (block j is a scalar 1), it makes nonzero contribution to only four blocks of matrix H, namely block (i,i), (i,j), (j,i) and (j,j). Matrix H is produced by the sum of these sparse matrices:

$$H = \sum_{i,j} \boldsymbol{J}_{ij}^\top(\mathbf{X}) \boldsymbol{J}_{ij}(\mathbf{X}). \tag{11}$$

According to the sparsity of matrix H, it can be divided into four blocks:

$$H = \begin{bmatrix} H_{\xi\xi} & H_{\xi M} \\ H_{M\xi} & H_{MM} \end{bmatrix}, \quad \text{where} \quad H_{M\xi} = H_{\xi M}^\top. \tag{12}$$

$H_{\xi\xi}$ is block diagonal and H_{MM} is diagonal. Since the linear system (9) is sparse, Schur complement [24] can be used to efficiently solve linear system (9). If we write step changes of pose parameters and intensities of voxels separately, (9) can be rewritten as follows:

$$\begin{bmatrix} H_{\xi\xi} & H_{\xi M} \\ H_{M\xi} & H_{MM} \end{bmatrix} \begin{bmatrix} \Delta\mathbf{X}_\xi \\ \Delta\mathbf{X}_M \end{bmatrix} = \begin{bmatrix} b_\xi \\ b_M \end{bmatrix}. \tag{13}$$

Using the Schur complement, we can compute step changes of poses and intensities of voxels sequentially from:

$$\left(H_{\xi\xi} - H_{\xi M} H_{MM}^{-1} H_{\xi M}^\top\right) \Delta\mathbf{X}_\xi = (b_\xi - H_{\xi M} H_{MM}^{-1} b_M), \tag{14}$$

$$H_{MM} \Delta\mathbf{X}_M = b_M - H_{\xi M}^\top \Delta\mathbf{X}_\xi. \tag{15}$$

Since matrix H_{MM} is diagonal, its inverse can be easily computed by the inverse of each element on the diagonal.

Algorithm 1: Summary of Direct Bundle Adjustment (DBA) to 3D Images

Input: A series of 3D images $\{I_i\}$.
Output: Global poses $\mathbf{X}_\xi$ and intensities of global image $\mathbf{X}_M$.

Step 1: Pre-computation:
Define volume M with zero intensities for all voxels;
Calculate gradient space of intensity of 3D images $\{I_1, I_2, ..., I_m\}$;
Interpolate intensity/gradient spaces of 3D images;
Step 2: Optimization:
Initialize the pose parameters $\mathbf{X}_\xi \leftarrow \mathbf{X}_{\xi_0}$;
Initialize M with initial pose $\mathbf{X}_{\xi_0}$ from local images;
while *Algorithm not converged* **do**
  for *every voxel* p_j *in volume* M **do**
    $p_{ij} = \mathbf{T}(\xi_i)p_j$;
    if $\sigma(p_{ij}) = 1$ **then**
      Calculate intensity difference e_{ij} by (3)
      Calculate Jacobian J_{ij} by (7);
    end
  end
  $J(\mathbf{X}) = [..., J_{ij}(\mathbf{X})^\top, ...]^\top$, $e(\mathbf{X}) = [..., e_{ij}, ...]^\top$;
  Calculate $\Delta\mathbf{X}_\xi$ by (14), calculate $\Delta\mathbf{X}_M$ by (15) ;
  Update poses, $\mathbf{T}(\mathbf{X}_\xi) \leftarrow \mathbf{T}(\Delta\mathbf{X}_\xi)\mathbf{T}(\mathbf{X}_\xi)$;
  Update intensities of M, $\mathbf{X}_M \leftarrow \mathbf{X}_M + \Delta\mathbf{X}_M$;
end

Solving (14) then (15) is much more efficient than solving (9) directly.

Pre-computation of Gradient and Intensity Space: According to (5), calculation of partial derivative $\partial I_i/\partial f$ is a computationally costly step in the process of GN iterations. Since $\partial I_i/\partial f = \partial I_i/\partial p_{ij}$ where $\partial I_i/\partial p_{ij}$ is the partial derivative of I_i w.r.t the corresponding coordinates of point p_{ij} on the local 3D image, we can pre-compute the gradient space of intensity of all the local 3D images before the GN iterations. And then, the gradient value at any point in local frames can be read directly from the gradient space during GN iterations. This strategy can greatly improve the effectiveness of computation. In addition, since p_{ij} are usually not integers, the intensities and gradient values of voxels are interpolated to reduce the error of pose estimation, which are also processed before the iterations.

The DBA algorithm is summarized in Algorithm 1. Poses are initialized using visual odometry (VO) estimated by direct method [9].

III. Experiments and Results

In this section, we use 3D TEE image data as an example and perform simulated and in-vivo experiments to validate the proposed DBA algorithm.

A. Simulated experiments

Since currently, there is no BA algorithm available for 3D image registration/fusion, sequential method registering two images at each time is commonly used if there are more than two images. In this section, we access the robustness and accuracy of the proposed DBA method by comparing it with sequential fusion method.

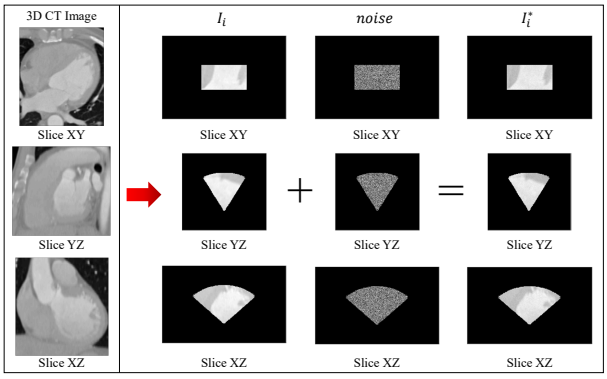

Fig. 3. Generation of the simulated 3D TEE images (3D images are displayed via 2D slices).

Pipeline of sequential fusion method to compare: In the sequential fusion experiments, we adopt the SSD as similarity measures for 3D image pairwise registration. The global image is initialized by the first image. Every time we designate the global image as a reference image and a new coming image as the moving image. Relative pose from moving frame to the global frame is calculated by the pairwise registration method. After registering the moving image to the global image, two images are fused together by averaging the intensities of voxels in the overlapping area. Then, the fused image is used as the global image and the reference image for the next coming image. The same processes are performed repeatedly until all images are fused.

Image data generation: We generate the simulated 3D TEE image data (in grayscale ranged from 0-255) by firstly cropping images from a 3D heart CT scan, and then adding intensity noises on them, which is illustrated in Fig. 3. In the first step, the corresponding image areas are cropped by using an empty 'viewfinder' which has the same FoV as a 3D TEE image. The 'viewfinder' is rotated and/or translated to the same frame as the 3D heart CT scan and the corresponding area of the image is cropped by the 'viewfinder'. In addition, it is well-known that speckle noise is an inherent property of medical ultrasound imaging [25], which generally results in the change of intensity of images, thereby influences the robustness of the direct method. Speckle noise after logarithmic compression can be assumed as White Gaussian Noise (WGN) [25], [26]. Therefore, WGN is added to the ground truth images obtained in the first step. Different levels of noises are generated by controlling the standard deviation (std) of the WGN.

1) Accuracy assessment: Five sequences of 3D TEE images are generated from different poses of 'viewfinder' and then WGN with a standard deviation of 8 is added to these images. Each sequence contains 11 simulated images and the ground truth pose of each image is listed in Table I (from Sequence 1 to Sequence 5).

We perform the experiments on these five sequences of image data by using the proposed DBA method and the sequential fusion method, respectively. In each experiment, we treat the first frame as the global frame. Therefore, 10

TABLE I

GROUND TRUTH POSES OF SIX SEQUENCES OF SIMULATED IMAGES

Sequence	Pose Parameter	Image ID in Each Group										
		I_1	I_2	I_3	I_4	I_5	I_6	I_7	I_8	I_9	I_{10}	I_{11}
1	E.A. (degree)	0	6	6	9	9	15	15	18	18	21	21
	Trans. (pixel)	0	5	25	5	25	5	25	5	25	5	25
2	E.A. (degree)	0	3	3	9	9	12	12	15	15	18	18
	Trans. (pixel)	0	15	25	15	25	15	25	15	25	15	25
3	E.A. (degree)	0	9	9	12	12	15	15	18	18	21	21
	Trans. (pixel)	0	15	20	15	20	15	20	15	20	15	20
4	E.A. (degree)	0	9	9	12	12	15	15	21	21	24	24
	Trans. (pixel)	0	10	25	10	25	10	25	10	25	10	25
5	E.A. (degree)	0	12	12	15	15	18	18	21	21	24	24
	Trans. (pixel)	0	5	15	5	15	5	15	5	15	5	15
6	E.A. (degree)	0	6	6	9	9	12	12	15	15	18	18
	Trans. (pixel)	0	10	20	10	20	10	20	10	20	10	20

E.A.: Euler angles, rotation order is X-Y-Z.
Trans.: Translations in X-Y-Z direction.
For example: for Sequence 1, image I_2, Euler angles are $[6,\ 6,\ 6]^\top$ degrees, translations are $[5,\ 5,\ 5]^\top$ pixels.
One-pixel length represents 0.25 mm in practice.

poses need to be estimated in each sequence.

Assessment of pose accuracy: In Fig. 4, mean absolute errors (MAE) of translation and Euler angles are compared respectively. From the results, it is shown clearly that the proposed DBA method has better accuracy than the sequential method. For the translation, results from the proposed DBA method have an accuracy of 0.1 pixels (0.025 mm in practice) for most of the cases. Only 3 out of 50 estimated translation errors are larger than 0.1 pixels. On the contrary, 37 out of 50 results from the sequential fusion method have errors larger than 0.1 pixels. For the accuracy of rotation, although both the sequential method and the proposed DBA method can obtain the results with high accuracy of 10^{-4} order of magnitude radians. The results from the proposed DBA method are still more accurate than the sequential fusion method for most of the cases.

Assessment of intensity accuracy: We also calculate the MAE of intensities of the fuse global images w.r.t. the ground truth image and the results are shown in the upper sub-table in Table II. In the five simulated experiments, The MAE of intensities of the fused image obtained from the proposed DBA method is always smaller than that from the sequential fusion method, which indicates the intensities of the fused image optimized from the proposed DBA method has better accuracy than those from the sequential method.

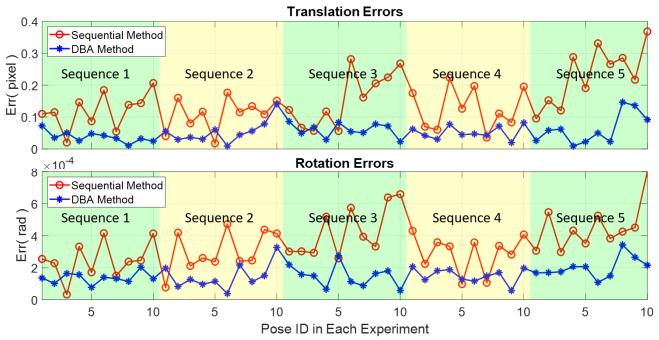

Fig. 4. Accuracy of the sequential fusion method and the proposed DBA method using the data from the Sequence 1 to Sequence 5. (Remark: One-pixel length represents 0.25 mm in practice.)

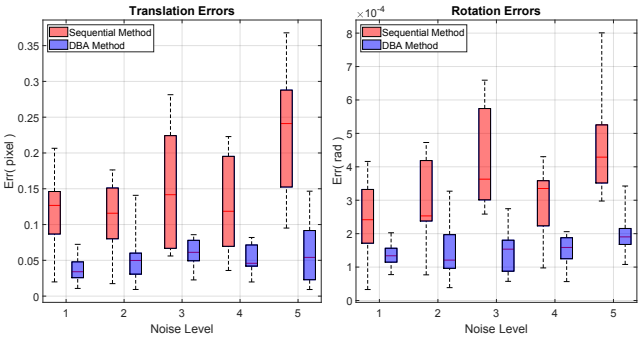

Fig. 5. Robustness of the sequential fusion method and the proposed DBA method using Sequence 6 with different intensity noise levels. (Remark: One-pixel length represents 0.25 mm in practice.)

TABLE II

COMPARISON OF INTENSITY ERRORS BETWEEN TWO METHODS

Sequence ID	1	2	3	4	5
Sequential	2.4833	2.7201	2.8225	2.6584	2.7031
DBA	2.1732	2.3689	2.4577	2.3613	2.3692
Sequence 6 (5 noise levels)	std 6 (WGN)	std 7 (WGN)	std 8 (WGN)	std 9 (WGN)	std 10 (WGN)
Sequential	1.9603	2.2835	2.6096	2.9332	3.2622
DBA	1.6822	1.9588	2.2364	2.5141	2.7923

2) Robustness assessment: Intensity noise is usually challenging for most of the direct methods. In this section, we assess the robustness of the proposed algorithm with different levels of intensity noises. The comparative experiments are performed using the sequential fusion method as well. The intensity noises are added based on Sequence 6 in Table I. We design five noise levels (Level 1 to Level 5) with the corresponding standard deviations of WGN as 6, 7, 8, 9, and 10 respectively.

Robustness of pose: The distribution of the MAE of translation and rotation is shown in Fig. 5. With the increase of the intensity noise from Level 1 to Level 5, the results of translation from the proposed DBA method can keep an accuracy of 0.1 pixels (0.025 mm in practice) for most of the cases, the largest error is around 0.15 pixel (0.03 mm in practice). By contrast, the error of translation from the sequential fusion method increases with fluctuation from noise Level 1 to Level 5 and the largest translation error is larger than 0.35 pixels. The MAE of rotation shows similar trends as that of translation. From the comparison of pose errors, it is shown that the proposed DBA algorithm has much better robustness to the intensity noise than the sequential fusion method although both methods can obtain a high accuracy.

Robustness of intensity: The intensity MAE of the reconstructed global image using two methods are compared and results are shown in the bottom sub-table in Table II. It is shown that although the MAE of intensities obtained by both methods increases with the increase of noise level, the results of the proposed DBA method is always better than that of the sequential fusion method.

B. In-vivo experiments

For the in-vivo experiments, datasets of 3D TEE images from five different patients are collected using a 2D array

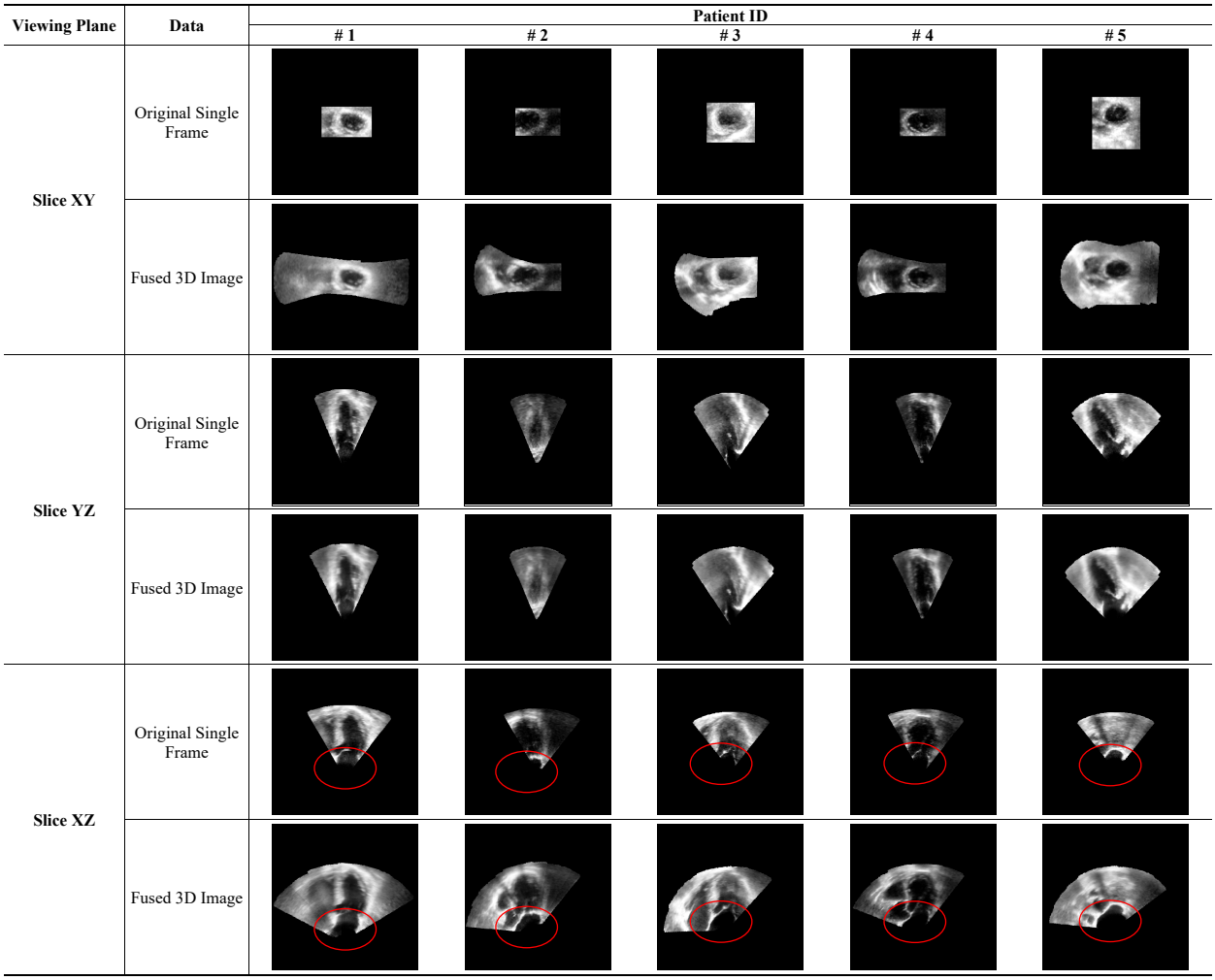

Fig. 6. Comparison of original single frame of in-vivo 3D TEE image with fused 3D global image using the proposed DBA algorithm. In the last two rows, regions with sharp boundaries are marked by circles for qualitative evaluation of the fused global images. (3D images are displayed via 2D slices.)

TABLE III
DESCRIPTION OF IN-VIVO DATASETS

Patient ID	No. frames	Volume size (voxel)	Resolution (mm/voxel)	FoV w.r.t original
# 1	9	240×160×208	0.69×0.98×0.73	2.10
# 2	6	240×160×208	0.77×1.11×0.82	2.02
# 3	6	277×208×208	0.58×0.87×0.63	2.01
# 4	6	240×160×208	0.64×0.92×0.68	1.80
# 5	8	277×208×208	0.63×0.90×0.68	2.06

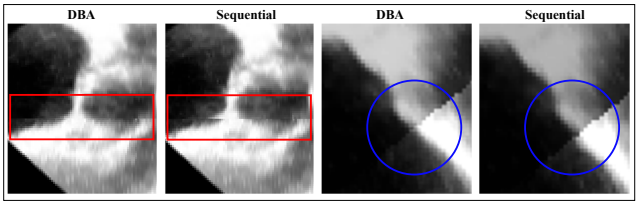

Fig. 7. Comparison of the aligned images using poses calculated by DBA and sequential method.

transducer with the assistance of the ECG-gating technique. ECG-gating is commonly used in the current 3D TEE imaging system to help capture 3D images of the heart at the same phase of different cardiac cycles [27]. Therefore, we can consider the registration problem as rigid. The details of the datasets are listed in the first four columns of Table III. In the in-vivo experiments, pose parameters are initialized using the VO estimated by direct method with similarity metric SSD and the global image is initialized by the local 3D images using the initial poses.

Evaluation of the results are performed based on the quality of the global images and aligned images since the ground-truth of in-vivo datesets are usually not available in practice. By comparing sharp boundaries in the single frame of images and fused images, such as left atrium (LA) walls indicated by circles in Fig. 6, it is found that the global images obtained from the proposed DBA have smooth transition and no misalignment/ghosting is found, which suggests good quality of the estimated global images. From the comparison in Fig. 6, it is evident that the FoV of the fused images are enlarged apparently. By counting the number of voxels, it is found that the FoV of the fused global images is enlarged to 2.10, 2.02, 2.01, 1.80, and 2.06 times as compared with the original single frame of TEE image of # 1, # 2, # 3, # 4, and # 5 respectively, which are listed in the last column of Table III. The results shows good potential of the proposed DBA for overcoming the drawback of limited FoV of 3D TEE images.

Furthermore, to access the accuracy of the pose estimation, images are aligned with the poses calculated by DBA method and sequential method, respectively. Two examples are shown in Fig. 7. We can find that the result of DBA has smooth transition, while the result of sequential method has apparent misalignments. In addition, mean squared error (MSE) between global images and registered images (transforming local images to the global frames by using the estimated poses) is calculated for the all five in-vivo datasets. The MSEs are $(235, 293, 244, 214, 321)$ and $(254, 311, 246, 233, 341)$, for DBA and sequential method, respectively. Smaller MSEs of DBA than sequential method suggest that the results of DBA method have better alignment than those of sequential method. Thus, all these results indicate that the proposed DBA outperforms the sequential method in terms of the pose accuracy.

IV. Conclusion

This paper proposes a novel direct bundle adjustment algorithm for 3D image fusion, which optimizes the intensities of the global image and the poses of the local frames simultaneously. Simulated and in-vivo experiments are performed to demonstrate that the proposed algorithm is robust to intensity noises and can obtain more accurate results than the conventional sequential fusion method. Experimental results using in-vivo 3D Transesophageal Echocardiography datasets from five different patients show that the fused 3D TEE images have around twice the field of view than the original image, indicating a significant potential clinical value of the proposed algorithm.

This new algorithm provides an elegant way to obtain the optimal global image and optimal local frame poses in one go, without any information loss or information reuse. Therefore, it can serve as a good framework for 3D image fusion. The current algorithm is still computationally expensive due to the large-scale optimization problem involved. In the future, different strategies will be investigated to significantly reduce the computational cost without sacrificing much on the quality of the results.

References

[1] H. Li and R. Hartley, "The 3d-3d registration problem revisited," in *2007 IEEE 11th international conference on computer vision*. IEEE, 2007, pp. 1–8.

[2] F. P. Oliveira and J. M. R. Tavares, "Medical image registration: a review," *Computer methods in biomechanics and biomedical engineering*, vol. 17, no. 2, pp. 73–93, 2014.

[3] K. Rajpoot, J. A. Noble, V. Grau, C. Szmigielski, and H. Becher, "Multiview rt3d echocardiography image fusion," in *International Conference on Functional Imaging and Modeling of the Heart*. Springer, 2009, pp. 134–143.

[4] L. Zhang, M. Ye, P. Giataganas, M. Hughes, and G.-Z. Yang, "Autonomous scanning for endomicroscopic mosaicing and 3d fusion," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2017, pp. 3587–3593.

[5] A. P.-W. Lee, Y.-Y. Lam, G. W.-K. Yip, R. M. Lang, Q. Zhang, and C.-M. Yu, "Role of real time three-dimensional transesophageal echocardiography in guidance of interventional procedures in cardiology," *Heart*, vol. 96, no. 18, pp. 1485–1493, 2010.

[6] M. A. Viergever, J. A. Maintz, S. Klein, K. Murphy, M. Staring, and J. P. Pluim, "A survey of medical image registration–under review," 2016.

[7] D. L. Hill, P. G. Batchelor, M. Holden, and D. J. Hawkes, "Medical image registration," *Physics in medicine & biology*, vol. 46, no. 3, p. R1, 2001.

[8] R. Szeliski and H.-Y. Shum, "Creating full view panoramic image mosaics and environment maps," in *Proceedings of the 24th annual conference on Computer graphics and interactive techniques*, 1997, pp. 251–258.

[9] Z. Mao, L. Zhao, S. Huang, Y. Fan, and A. P.-W. Lee, "Direct 3d ultrasound fusion for transesophageal echocardiography," *Computers in Biology and Medicine*, vol. 134, p. 104502, 2021.

[10] V. Grau, H. Becher, and J. A. Noble, "Registration of multiview real-time 3-d echocardiographic sequences," *IEEE transactions on medical imaging*, vol. 26, no. 9, pp. 1154–1165, 2007.

[11] M. C. Carminati, C. Piazzese, L. Weinert, W. Tsang, G. Tamborini, M. Pepi, R. M. Lang, and E. G. Caiani, "Reconstruction of the descending thoracic aorta by multiview compounding of 3-d transesophageal echocardiographic aortic data sets for improved examination and quantification of atheroma burden," *Ultrasound in Medicine & Biology*, vol. 41, no. 5, pp. 1263–1276, 2015.

[12] H. W. Mulder, M. van Stralen, B. Ren, A. Haak, M. A. Viergever, J. G. Bosch, and J. P. Pluim, "Atlas-based mosaicing of left atrial 3-d transesophageal echocardiography images," *Ultrasound in medicine & biology*, vol. 43, no. 4, pp. 765–774, 2017.

[13] C. Wachinger, W. Wein, and N. Navab, "Three-dimensional ultrasound mosaicing," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2007, pp. 327–335.

[14] E. G. Learned-Miller, "Data driven image models through continuous joint alignment," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 2, pp. 236–250, 2005.

[15] M. Cox, S. Sridharan, S. Lucey, and J. Cohn, "Least-squares congealing for large numbers of images," in *2009 IEEE 12th International Conference on Computer Vision*. IEEE, 2009, pp. 1949–1956.

[16] L. Zöllei, E. Learned-Miller, E. Grimson, and W. Wells, "Efficient population registration of 3d data," in *International Workshop on Computer Vision for Biomedical Image Applications*. Springer, 2005, pp. 291–301.

[17] B. Triggs, P. F. McLauchlan, R. I. Hartley, and A. W. Fitzgibbon, "Bundle adjustment—a modern synthesis," in *International workshop on vision algorithms*. Springer, 1999, pp. 298–372.

[18] A. Delaunoy and M. Pollefeys, "Photometric bundle adjustment for dense multi-view 3d modeling," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1486–1493.

[19] H. Alismail, B. Browning, and S. Lucey, "Photometric bundle adjustment for vision-based slam," in *Asian Conference on Computer Vision*. Springer, 2016, pp. 324–341.

[20] M. Zollhöfer, A. Dai, M. Innmann, C. Wu, M. Stamminger, C. Theobalt, and M. Nießner, "Shading-based refinement on volumetric signed distance functions," *ACM Transactions on Graphics (TOG)*, vol. 34, no. 4, pp. 1–14, 2015.

[21] R. Maier, K. Kim, D. Cremers, J. Kautz, and M. Nießner, "Intrinsic3d: High-quality 3d reconstruction by joint appearance and geometry optimization with spatially-varying lighting," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 3114–3122.

[22] J. Engel, V. Koltun, and D. Cremers, "Direct sparse odometry," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 3, pp. 611–625, 2017.

[23] B. Hall, *Lie groups, Lie algebras, and representations: an elementary introduction*. Springer, 2015, vol. 222.

[24] F. Zhang, *The Schur complement and its applications*. Springer Science & Business Media, 2006, vol. 4.

[25] C. P. Loizou, C. S. Pattichis, C. I. Christodoulou, R. S. Istepanian, M. Pantziaris, and A. Nicolaides, "Comparative evaluation of despeckle filtering in ultrasound imaging of the carotid artery," *IEEE transactions on ultrasonics, ferroelectrics, and frequency control*, vol. 52, no. 10, pp. 1653–1669, 2005.

[26] J. Zhang and Y. Cheng, *Despeckling Methods for Medical Ultrasound Images*. Springer, 2020.

[27] R. M. Lang, L. P. Badano, W. Tsang, D. H. Adams, E. Agricola, T. Buck, F. F. Faletra, A. Franke, J. Hung, L. P. de Isla, *et al.*, "Eae/ase recommendations for image acquisition and display using three-dimensional echocardiography," *European Heart Journal–Cardiovascular Imaging*, vol. 13, no. 1, pp. 1–46, 2012.