Medical Image Analysis 84 (2023) 102709

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Robust endoscopic image mosaicking via fusion of multimodal estimation

Liang Li [a,e,*], Evangelos Mazomenos [a], James H. Chandler [b], Keith L. Obstein [c,d], Pietro Valdastri [b], Danail Stoyanov [a], Francisco Vasconcelos [a]

[a] Wellcome/EPSRC Centre for Interventional and Surgical Sciences(WEISS) and Department of Computer Science, University College London, London, UK
[b] Storm Lab UK, School of Electronic, and Electrical Engineering, University of Leeds, Leeds LS2 9JT, UK
[c] Division of Gastroenterology, Hepatology, and Nutrition, Vanderbilt University Medical Center, Nashville, TN 37232, USA
[d] STORM Lab, Department of Mechanical Engineering, Vanderbilt University, Nashville, TN 37235, USA
[e] College of Control Science and Engineering, Zhejiang University, Hangzhou, 310027, China

ARTICLE INFO

Keywords:
Medical image processing
Optical flow
Image mosaicking
Pose graph optimisation
Endoscopic image mosaicking

ABSTRACT

We propose an endoscopic image mosaicking algorithm that is robust to light conditioning changes, specular reflections, and feature-less scenes. These conditions are especially common in minimally invasive surgery where the light source moves with the camera to dynamically illuminate close range scenes. This makes it difficult for a single image registration method to robustly track camera motion and then generate consistent mosaics of the expanded surgical scene across different and heterogeneous environments. Instead of relying on one specialised feature extractor or image registration method, we propose to fuse different image registration algorithms according to their uncertainties, formulating the problem as affine pose graph optimisation. This allows to combine landmarks, dense intensity registration, and learning-based approaches in a single framework. To demonstrate our application we consider deep learning-based optical flow, hand-crafted features, and intensity-based registration, however, the framework is general and could take as input other sources of motion estimation, including other sensor modalities. We validate the performance of our approach on three datasets with very different characteristics to highlighting its generalisability, demonstrating the advantages of our proposed fusion framework. While each individual registration algorithm eventually fails drastically on certain surgical scenes, the fusion approach flexibly determines which algorithms to use and in which proportion to more robustly obtain consistent mosaics.

1. Introduction

Image mosaicking, or image stitching, is an established technique in computer vision that is now widely utilised in robotics and consumer products such as cell phones. In minimally invasive surgeries guided by a camera scope with a narrow field of view, mosaicking can generate an expanded view of operative site that can aid the surgeon in navigating instruments and planning the surgery (Kutarnia and Pedersen, 2015). Unlike very well established mosaicking applications involving indoor or outdoors scenes (Oliveira et al., 2015; Xu et al., 2020; Chon et al., 2007), mosaicking of endoscopic images has significantly increased challenges (Loewke et al., 2020; Richa et al., 2014; Loewke et al., 2010) that can take multiple forms that we now detail. The scene illumination is severely non-homogeneous as the only light source comes from the endoscopic camera itself and moves within the environment. Tissue and organs are very prone to saturated specular reflections that dynamically change with the camera motion (Wu and Su, 2017). Tissue can be dynamically occluded by blood or other artefacts (e.g. floating particles in fetoscopy) that have motion patterns inconsistent with the camera motion (Reeff et al., 2006). There are non-rigid tissue deformations caused by breathing, blood flow, or surgical instrument manipulation (Zhou and Jayender, 2021). The entire visualisation of the operative site can take a significant amount of time within the surgical workflow, and long-term mosaic consistency cannot be ignored (Li et al., 2021). Finally, the visual appearance of different environments/organs vary significantly, making feature extraction difficult to generalise. An algorithm that may work robustly in a narrowly defined environment will eventually degrade or fail when there are substantial changes in scene appearance (Bano et al., 2020). Addressing these challenges in a robust way is fundamental since an image mosaic can be rendered unusable with just a short number of poorly estimated image registrations.

* Corresponding author.
E-mail addresses: liang.li@zju.edu.cn (L. Li), e.mazomenos@ucl.ac.uk (E. Mazomenos), J.H.Chandler@leeds.ac.uk (J.H. Chandler), keith.obstein@vanderbilt.edu (K.L. Obstein), p.valdastri@leeds.ac.uk (P. Valdastri), danail.stoyanov@ucl.ac.uk (D. Stoyanov), f.vasconcelos@ucl.ac.uk (F. Vasconcelos).

https://doi.org/10.1016/j.media.2022.102709
Received 8 November 2021; Received in revised form 15 August 2022; Accepted 29 November 2022
Available online 14 December 2022

Generally, image mosaicking consists of different sub-problems. The first one is the data association between common parts of the scene under different views (Huang and Netravali, 2002). The second is the estimation of a geometric transformation that is consistent with the data association and maps different views into a single mosaic image (Chum and Matas, 2012). These two sub-problems also be tackled simultaneously (*e.g.*, direct registration (Bartoli, 2008)). Finally, the image intensities of individual images need to be blended in a consistent and smooth mosaic (Tian and Shi, 2014). One can also consider the global optimisation of long-term mosaics as a separate sub-problem (Zhou and Jayender, 2021; Li et al., 2021). The first sub-problem of data association is the most challenging in surgical scenes and draws a significant amount of research attention. The most classic approach is to detect and extract image point features corresponding to unique landmarks in the scene and then match them across different views. This feature-based mosaicking approach (Milgram, 1975) has been investigated extensively in recent decades, using different well-known hand-crafted feature approaches such as Harris (Okumura et al., 2013), SIFT (Li et al., 2008), SURF (Rong et al., 2009), ORB (Chaudhari et al., 2017), and FAST (Wang et al., 2012). More recently, data-driven features that are learned by deep neural networks have been utilised for image mosaicking (Bano et al., 2020; Zhang et al., 2019). There are also mosaicking approaches that do not rely on feature extraction. Direct and dense pixel-based registration methods can be formulated as an iterative optimisation problem by maximising the similarity computed with mutual information (Miranda-Luna et al., 2008) or other photometric similarity/difference metrics (Levin et al., 2004; Konen et al., 2007). With the popularisation of deep learning in different problems, some end-to-end mosaicking algorithms based on deep learning regression of registration parameters (Bano et al., 2019; Nguyen et al., 2018) have been proposed.

There is also research focused on developing image mosaicking methods that are dedicated to deal with surgical scenes and its associated challenges. In Zhou and Jayender (2021), non-rigid Simultaneous Localization And Mapping (SLAM) is adopted to account for tissue deformation. Similarly in Loewke et al. (2010), deformation and cumulative errors are addressed with local and global alignment. In Soper et al. (2012), structure-from-motion with bundle adjustment is utilised to reduce cumulative errors when generating the mosaic. In Gong et al. (2021), the non-rigid deformation is estimated with a parametric free form deformation model. By reviewing the literature, it is clear that most of the existing mosaicking algorithms in this domain focus on estimating the deformation or reducing cumulative errors. However, the problems of inconsistent light conditions and environment changes have not been analysed in detail. These can happen frequently when generalising a method to work robustly on different cases, where the camera scope and light source may have settings, or a different patient may have anatomy structures with different appearance. With these challenges in mind, this paper aims to solve the problem of robustness in image data association for mosaicking of surgical scenes. Instead of choosing between point feature extraction, optimisation of photometric alignment, or a deep learning approach, we propose to fuse multi-modal estimation to bring the best of each method. Our fusion framework is agnostic to the data source and can be easily generalised to other contexts. In this paper we consider as an exemplary case the fusion of three sources: optical flow, hand-crafted (scale-invariant feature transform) SIFT features, and direct photometric registration. The considered optical flow method is the end-to-end deep neural network FlowNet2.0 (Ilg et al., 2017). After data association between different frames, the geometric alignment between different views is modelled as a homography linear mapping, and approximated as an affine transformation. For both hand-crafted features and optical flow, (random sample consensus) RANSAC is used to filter out outliers prior to registration estimation. The core of our proposed method is to take all the available and competing motion estimation approaches as inputs to a pose graph optimisation framework. Considering different camera views as graph nodes, up to three edges representing different motion estimations will link them. The optimal graph state is computed using the Levenberg–Marquardt (L–M) algorithm on the affine Lie group. The experimental results show that the proposed fusion-based image mosaicking algorithm outperforms keypoint feature-based, dense registration, and end-to-end algorithms in terms of robustness, consistency and generalisation to different datasets. Therefore, the contributions of this paper are threefold:

1. We propose a framework to fuse different image data association algorithms based on their uncertainties for endoscopic mosaicking. The proposed method improves robustness and adaptability across various types of surgical scenes.
2. The proposed fusion scheme is formulated in general form and is not constrained to any specific estimation sources, nor to the type of surgery. It can easily be extended to other problems involving multi-modal estimation and/or data sources.
3. Extensive experiments in significantly different surgeries are carried out to validate the generalisability of the proposed method. We test more than ten sequences of ex-vivo laparoscopic video from the publicly available SCARED dataset (Allan et al., 2021), a publicly available fetoscopic surgery dataset (Bano et al., 2019), and also cadaver sequences captured with the Bellowscope robotic gastric endoscopy platform (Chandler et al., 2020; Garbin et al., 2018, 2019). The fusion approach is compared against the individual estimation approaches, *i.e.*, SIFT-based, direct registration-based, end-to-end deep learning-based mosaicking.

The remainder of this paper is organised as follows. Section 2 gives a review of the related work. Section 3 introduces the formulations in correspondence matching and homography estimation, and details our proposed fusion-based mosaicking framework. Section 4 presents and discusses the experimental results. We finally conclude the paper and provide some remarks on future work in Section 5.

2. Related work

While image mosaicking is a problem with a wide variety of well established application domains, medical imaging has its own dedicated challenges. Therefore, in this section we concentrate on methods directly applied to surgical data. The algorithms can be classified into three categories: feature-based, direct, and deep learning-based.

Feature-based mosaicking has been studied for decades in the context of medical imaging. Early work can be found in Can et al. (2002b,a), where the edges of vascular centrelines in human retina are used as features. To speed up the mosaicking, a hierarchical registration algorithm was adopted. In Lee and Bajcsy (2005), the centroids of vascular regions were selected as features for image registration and mosaicking, and a normalised correlation-based registration algorithm is used to estimate affine transformations. In Bergen et al. (2009), the authors used corner-like features and Kanade–Lucas–Tomasi tracker (KLT) to track features in subsequent frames rather than using feature matching. With the development of abstract feature extraction in the community of computer vision, some state-of-the-art feature descriptors were also utilised for mosaicking of medical images. Despite its relatively old age, SIFT (Lowe, 1999) is one of the most widely utilised feature descriptors (Daga et al., 2016; Richa et al., 2014; Jalili et al., 2020). Other popular handcrafted approaches include improved modifications of SIFT (Yu et al., 2015; Li et al., 2017; Gupta et al., 2016) and (speeded up robust features) SURF (Bay et al., 2006; Reeff et al., 2006). (oriented FAST and rotated BRIEF) ORB features have also been utilised for mosaicking in the context of robotic endomicroscopy (Rosa et al., 2018). Notably, here the authors fuse robot movement information with image registration to produce more robust estimation. Independently of utilising different image feature descriptors, we also note that different feature matching algorithms can also be considered (Viergever et al., 2016; Sotiras et al., 2013).

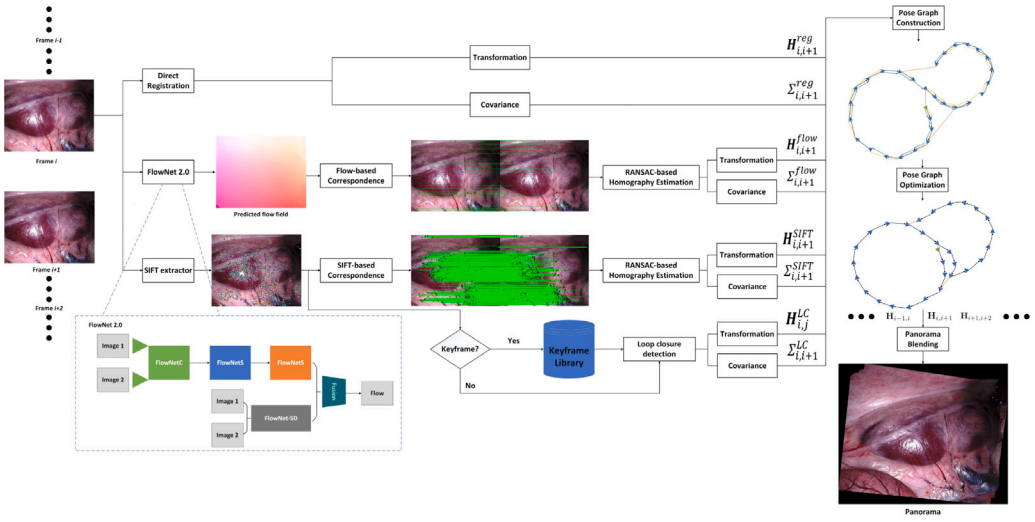


Fig. 1. The diagram of the proposed method. There are three component homography estimation algorithms, *i.e.*, SIFT-based, direct registration-based, and the optical flow-based. The pose graph is constructed based on the three estimation sources with their own uncertainties respectively. The optimal state is obtained by optimising the cost function in the affine Lie group. Finally, the panorama can be generated with the optimal homography matrices.

Unlike feature-based methods, direct registration aims at using the information of all image pixels (Baum et al., 2021). In Ogien et al. (2020), normalised cross-correlation (NCC) was used to maximise similarity between registered images. Furthermore, in Capek and Krekule (1999), three similarity-based methods were studied, the sum of absolute valued differences (SAVD), NCC, and the mutual information function (MIF). It was reported that SAVD performs best in terms of computational efficiency, NCC is more robust to uncorrelated stochastic noise, and MIF outperforms the other two in terms of nonlinear corruption of the intensity scales of the image. In Seshamani et al. (2009), the sum of squared differences (SSD) was used to measure difference of the two images in terms of their intensities. The first-order Taylor linearisation was utilised to minimise the SSD. Moreover, bundle adjustment that minimises the total point re-projection error was adopted for global optimisation. In Seshamani et al. (2006), the authors adopted a two-step optimisation algorithm. In the first stage, an initial estimate of 2D translation is computed by performing a brute-force search to maximise normalised cross-correlation between images; In the second stage, a local continuous finetune is applied by minimising intensity difference of the two images. In Peter et al. (2018), a pixel-wise image gradient alignment was adopted to highlight vessel-like structures. In Richa et al. (2014), an SSD-like function was used to minimise the intensity dissimilarities between two images. And a non-rigid illumination compensation fine-tuning was adopted to model local variations. Some studies try to combine feature-based and direct methods to take advantage of both (Richa et al., 2012).

Deep learning based methods have drawn more attention in recent years. Some researchers made efforts to utilise deep learning either in feature extraction or end-to-end transformation regression. In Bano et al. (2020), U-Net (Ronneberger et al., 2015) was used to segment vessels in fetoscopic images. Then, a direct image registration based on the output probability map of the neural network was proposed to estimate the homography matrix. In Bano et al. (2020), a deep image homography with controlled data augmentation was proposed to estimate homography between the two input images directly. To the best of the authors' knowledge, there is no study addressing the robustness and generalisation challenges across different environments in the context of surgical video mosaicking. This is an important problem that can arise in different domains, including GI endoscopy, fetoscopic surgery, and laparoscopy. This paper proposes to solve robustness challenges and generalisability by fusing multimodal estimation within an affine pose graph framework.

3. Approach

This section presents the proposed algorithm for surgical video mosaicking. The diagram of the proposed algorithm is presented in Fig. 1, displaying its several different components. It contains the three baseline data association methods (optical flow, handcrafted features, direct registration). Additionally, a loop closure detection source is included based on storing handcrafted features in keyframes. The fusion of multimodal results is achieved with pose graph optimisation, and finally image stitching and blending is performed based on the optimised graph. While direct registration performs data association and registration simultaneously, both handcrafted features and optical flow only perform the data association. Thus they require a second step to estimate homography transformations via 4-point linear estimation within a RANSAC robust estimator. We fuse all three methods together with pose graph optimisation to make the estimation more robust. Finally, all the images can be stitched together with respect to the middle frame within the sequence.

3.1. Optical flow-based correspondence

While both direct pixel-based registration and feature-based registration are classic approaches that have been extensively described in the previous works, image registration based on general optical flow is less common, especially with more recent deep learning methods, and therefore we provide here a more detailed account.

Optical flow measures displacement of pixels in two images. It is computed based on the assumption that intensity of the same object is constant in the consecutive frames, *i.e.*:

$$I(x, y, t) = I(x + dx, y + dy, t + dt) \tag{1}$$

where $I(x, y, t)$ denotes the intensity of pixel (x, y) at time t. Currently, there are several ways to compute the optical flow. The first type is the gradient-based method that includes Lucas-Kanade and Horn–Schunck. It assumes that the optical flow is smooth on the entire image. While in our case, specularities may degenerate the computation. The second type is the matching-based method that starts from sparse feature correspondences and interpolates a flow field for every pixel. This is not suitable in our case as it would be redundant with feature-based registration and fail in the same cases. The third type is the energy minimisation approach that uses the dense information of the

whole image. Again, this approach may be affected by inconsistent specular reflections. More recently, the energy minimisation methods can be optimised on training data with a deep neural network. In this paper, we use the deep learning-based method FlowNet2.0 due to its state-of-the-art performance. It has three types of components: FlowNetSimple (FlowNetS), FlowNetCorrelation (FlowNetC) which are proposed in FlowNet (Dosovitskiy et al., 2015), and FlowNet-Small-Displacement (FlowNet-SD) which is finetuned on a stack of FlowNetS and FlowNetC. Input of FlowNetS is a stack of two images, and the network architecture follows an encoder–decoder framework. It has six convolutional layers, four deconvolutional layers and finally a bilinear upsampling to lift the prediction map to full image resolution.

In order to make the network more efficient at catching salient features, FlowNetC first adopts two independent, yet identical processing streams for the two images separately. The two embeddings are then combined with a correlation layer that aids the following network to find correspondence. Given two patches centred at $\mathbf{x}_1$ in the first feature map $\mathbf{f}_1$ and $\mathbf{x}_2$ in the second feature map $\mathbf{f}_2$, the correlation layer is:

$$c(\mathbf{x}_1, x_2) = \sum_{\mathbf{o}\in[-k,k]\times[-k,k]} \langle \mathbf{f}_1(\mathbf{x}_1 + \mathbf{o}), \mathbf{f}_2(\mathbf{x}_2, \mathbf{o})\rangle \tag{2}$$

Note that it is identical to one-step convolution with the kernel as data from another feature map rather than the filter. It limits the maximal displacement k to within only the local neighbourhood to reduce the computation. After the correlation layer, FlowNetC adopts the FlowNetS to predict the optical flow. It was reported in Ilg et al. (2017) that FlowNetC outperforms FlowNetS if training under the same condition. FlowNet-SD is based on a stack of one FlowNetC and two FlowNetS (FlowNet-CSS). FlowNet-SD deepens the network with multiple layers with 3×3 kernels at the beginning of the network and is trained on dataset with small displacement. Finally, FlowNet2.0 fuses FlowNet-CSS and FlowNet-SD to give the predicted flow field to full resolution as the input image. One benefit of FlowNet2.0 is its generalisation, *i.e.*, we do not need to re-train it on surgical data for our mosaicking task. The network is trained from simple to more realistic datasets, *i.e.*, from the FlyingChairs synthetic dataset, to the FlyingThings3D synthetic dataset (Mayer et al., 2016), and finally on the KITTI real video dataset (Geiger et al., 2012). Examples of the predicted flow field on the endoscopic data are shown in Fig. 2(e).

The correspondence between the two images can be obtained by the flow field:

$$\begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} u' \\ v' \end{bmatrix} + \begin{bmatrix} o_x \\ o_y \end{bmatrix} \tag{3}$$

where $\begin{bmatrix} u & v \end{bmatrix}^\top$ is the position of the keypoint in the target image, and $\begin{bmatrix} u' & v' \end{bmatrix}^\top$ is the corresponding keypoint in the source image, $\begin{bmatrix} o_x & o_y \end{bmatrix}^\top$ is the value of optical flow. An example of the correspondence estimation based on the optical flow is shown in Fig. 2(e). From this point onwards, pairwise point correspondences between two frames are established and the remaining registration pipeline is identical to estimation with sparse feature correspondences, *i.e.*, SIFT in this paper.

3.2. Homography estimation

Both optical flow and SIFT provide pairwise point correspondences, dense and sparse respectively, and estimating a homography registration can be made identical for both methods. On the other hand, the computation of homographies based on direct pixel-based image registration is jointly done with data association. In this subsection, we first derive the correspondence-based homography estimation, then the direct registration-based homography estimation. The transformation between correspondence pairs can be modelled as:

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = s \cdot \mathbf{H} \cdot \begin{bmatrix} u' \\ v' \\ 1 \end{bmatrix} = s \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{bmatrix} \begin{bmatrix} u' \\ v' \\ 1 \end{bmatrix} \tag{4}$$

Table 1
Number of frames of the sequences for experiment.

Dataset	Sequence umber	Number of frames
SCARED	Seq. 1	196
	Seq. 2	279
	Seq. 3	87
	Seq. 4	447
	Seq. 5	347
Fetoscopy	Seq. 1	400
	Seq. 2	300
	Seq. 3	150
	Seq. 5	200
	Seq. 6	200
Human cadaver	Seq. 1	30
	Seq. 2	51
	Seq. 3	20
	Seq. 4	20
	Seq. 5	100

where $\mathbf{H}$ is a 3×3 homography matrix, *i.e.*, $h_{33} = 1$, and s is the scaling factor. In theory, the transformation is projective with a scaling factor. While as indicated in Bano et al. (2020), Peter et al. (2018), Li et al. (2021), approximating it with affine transformation gives more stable results for the endoscopic mosaicking. Thus, we follow this conclusion and assume $s \cdot \mathbf{H}$ to be affine, *i.e.*, $s = 1$, $h_{31} = h_{32} = 0$, $\begin{bmatrix} h_{11} & h_{12} \\ h_{21} & h_{22} \end{bmatrix}$ is an arbitary non-singular matrix, and $\begin{bmatrix} h_{13} \\ h_{23} \end{bmatrix}$ is the translation vector. The correspondence pair can be either from optical flow or SIFT. Every correspondence pair gives two constraints:

$$\begin{aligned} u &= \frac{h_{11}u' + h_{12}v' + h_{13}}{h_{31}u' + h_{32}v' + 1} \\ v &= \frac{h_{21}u' + h_{22}v' + h_{23}}{h_{31}u' + h_{32}v' + 1} \end{aligned} \tag{5}$$

If there are n pairs of correspondence, a linear matrix equation can be obtained:

$$\underbrace{\begin{bmatrix} u'_1 & v'_1 & 1 & 0 & 0 & 0 & -u_1u'_1 & -u_1v'_1 & -u_1 \\ 0 & 0 & 0 & u'_1 & v'_1 & 1 & -v_1u'_1 & -v_1v'_1 & -v_1 \\ & & & & \vdots & \vdots & & & \\ u'_n & v'_n & 1 & 0 & 0 & 0 & -u_nu'_n & -u_nv'_n & -u_n \\ 0 & 0 & 0 & u'_n & v'_n & 1 & -v_nu'_n & -v_nv'_n & -v_n \end{bmatrix}}_{\mathbf{A}_{2n\times9}} \underbrace{\begin{bmatrix} h_{11} \\ h_{12} \\ h_{13} \\ h_{21} \\ h_{22} \\ h_{23} \\ h_{31} \\ h_{32} \\ 1 \end{bmatrix}}_{\mathbf{h}_9} = \underbrace{\begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 0 \end{bmatrix}}_{\mathbf{0}_{2n}} \tag{6}$$

So, if we have four pairs of correspondence, Eq. (6) can be solved. If there are more than four pairs of correspondence, the optimal solution can be obtained by minimising

$$\arg\min_{\mathbf{h}} \|\mathbf{A}\mathbf{h} - \mathbf{0}\| = \arg\min_{\mathbf{h}} \|\mathbf{A}\mathbf{h}\| \tag{7}$$

As we have the result of optical flow for every pixel, we can have $w \times h$ pairs of correspondence in theory, where w, and h are the width and height of the image respectively. Even for the SIFT, there may be hundreds of thousands pairs of corresponding points. Solving the problem by minimising Eq. (7) has two problems: First, computation of the optimisation problem will be very high; Second, the outliers and noise of the correspondence estimation may deteriorate the estimated homography matrix. So in this paper, we use RANSAC with the 4-point method to identify inliers and outliers. Here, the recognition of inliers and outliers is based on the distance in pixels between a point in one image and its re-projected correspondence from the other image through the transformation $\hat{\mathbf{H}}$. And ε is the threshold set by the user to identify outliers in the flow field.

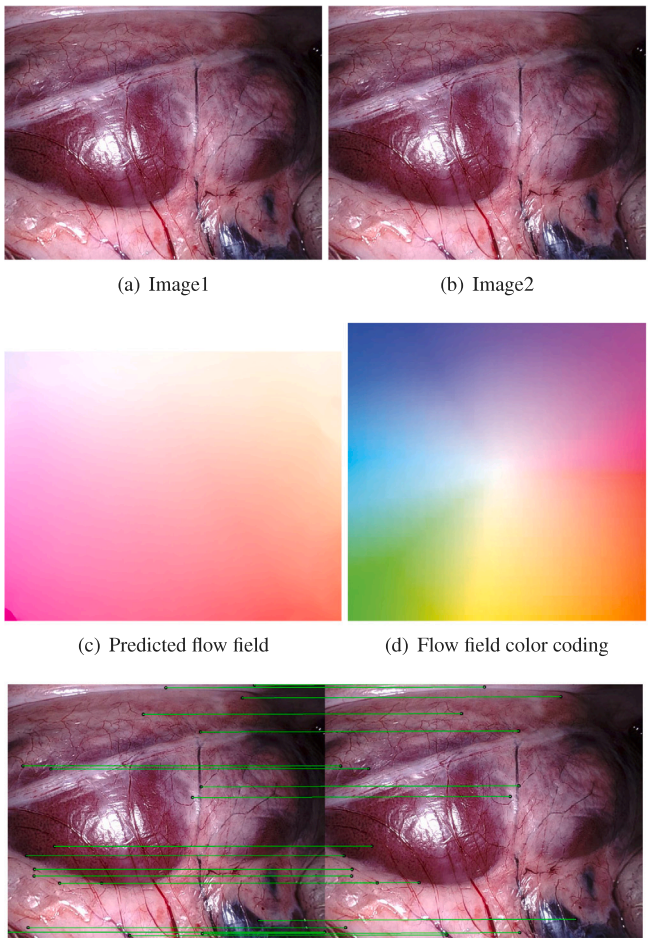


Fig. 2. An example of the results of optical flow prediction and correspondence establishment. (a) and (b) show the two input images, and (c) is the predicted flow field by the Flownet2.0, where the colour coding scheme is shown in (d). The correspondence can be established using Eq. (3). In theory, the correspondence is very dense as correspondence for most pixels can be computed except ones close to the image border. Only a small portion of the correspondence is presented in (e) for a better visualisation.

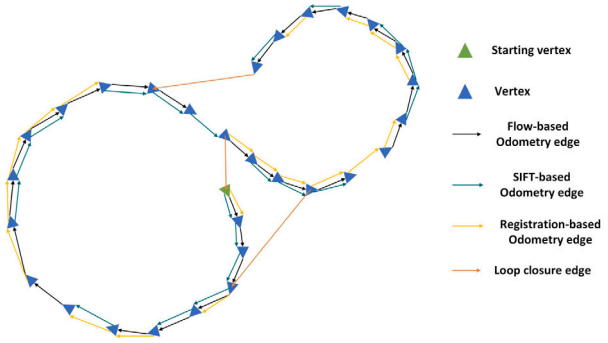


Fig. 3. An illustration of the pose graph that is constructed using the optical flow, SIFT, direct registration, and loop closure detection. The nodes are denoted in blue triangles. And the different types of edges are denoted in lines with different colours.

The homography matrix can be also estimated using direct image registration with the photometric loss:

$$\mathbf{L}(\mathbf{H}_{i,i+1}) = \|I_i - \mathbf{T}(I_{i+1}, \mathbf{H}_{i,i+1})\| \tag{8}$$

where the function $\mathbf{T}(I_{i+1}, \mathbf{H}_{i,i+1})$ warps the image I_{i+1} with transformation. By transforming the image I_{i+1} into a new position using $\mathbf{H}_{i,i+1}$, we

Fig. 4. Examples of mosaicking directly obtained from using the robot kinematics, extracted from seq. 1 (a) and seq. 5 (b) of the SCARED dataset. The kinematics are not accurate enough to generate mosaics.

can obtain I_{i+1} in its new position and view as $\tilde{I}_{i+1} = \mathbf{T}(I_{i+1}, \mathbf{H}_{i,i+1})$. The difference of Image I_i and $\tilde{I}_{i+1}$ can be computed with their L2 norm. The optimal transformation matrix can be obtained by minimising the loss:

$$\mathbf{H}^{reg}_{i,i+1} = \arg\min_{\mathbf{H}_{i,i+1}} \mathbf{L}(\mathbf{H}_{i,i+1}) \tag{9}$$

The optimisation of Eq. (9) is based on a standard pyramidal Lucas–Kanade registration framework that minimises the least-square difference (photometric loss) between a fixed frame I_i and a warped moving image $\tilde{I}_{i+1}$. This optimisation problem in Eq. (9) can be solved with the L–M iterative algorithm in an iterative way.

3.3. Fusion of multimodal estimations

For every pair of consecutive images, there are three possible estimated transformation matrices, *i.e.*, $\mathbf{H}^{flow}_{i,i+1}$, $\mathbf{H}^{SIFT}_{i,i+1}$, and $\mathbf{H}^{reg}_{i,i+1}$. A more robust estimation result can be obtained by fusing the three estimation sources. Inspired by the SLAM literature in mobile robotics, we perform the fusion via pose graph optimisation. The graph to be optimised can be constructed as $\mathcal{G} = \{\mathbf{V}, \mathbf{E}\}$, where $\mathbf{V} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$ is the set of vertices and $\mathbf{E} = \{\mathbf{z}_{1,2}, \mathbf{z}_{2,3}, \dots, \mathbf{z}^{lc}\}$ is the set of edges. Both $\mathbf{x}$ and $\mathbf{z}$ are affine matrices, and $\mathbf{z}$ are the estimated transformation matrices $\mathbf{H}^{flow}$, *etc*. An illustration of the pose graph is shown in Fig. 3. Any vertex $\mathbf{x}_i$ in the graph represents the transformation of the ith image with respect to the anchor (first) image, and they constitute the state to be estimated (optimised). The edges define constraints between pairs of vertices, which can be provided by the affine homography estimations obtained from optical flow, SIFT-correspondences, and direct registration.

Additionally, edges can also be loop closure constraints, i.e. registration of non-consecutive frames when a scene is revisited. Loop closure detection is based on SIFT keypoint features extracted from a set of key frames. The first frame in the sequence is always a key frame. If the current movement with respect to the latest key frame is larger than either a distance or a time threshold, the current frame is defined as a new key frame. SIFT features of the key frames are stored using bag-of-words. The similarity between a new frame and every other key frame will be computed to check if the camera revisits previous scenes. Every estimated transformation matrix is associated with a covariance matrix $\mathbf{\Sigma}_{i,i+1}$ representing how certain the estimation is. The covariance of the flow-based transformation $\mathbf{\Sigma}^{flow}_{i,i+1}$ is computed based on the ratio of inliers. The covariance of the SIFT-based transformation $\mathbf{\Sigma}^{SIFT}_{i,i+1}$ is jointly based on ratio of inliers and number of features. The covariance of the direct registration-based transformation $\mathbf{\Sigma}^{reg}_{i,i+1}$ is based on the finally minimal photometric loss. Then we can define the cost function

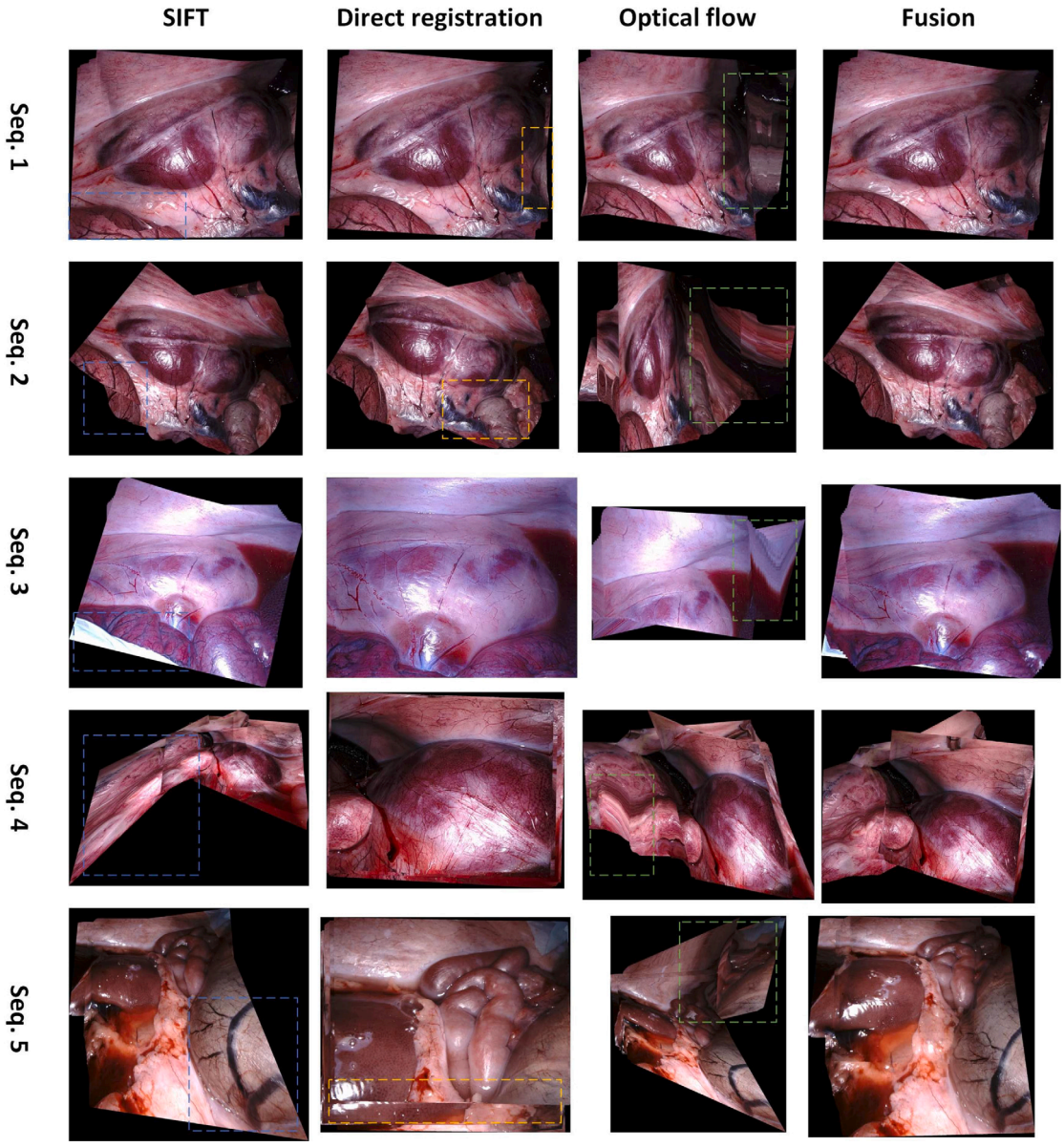


Fig. 5. Results on the SCARED dataset. Mosaicking results for five sequences are presented from the first to the last row. The SIFT, direct registration, optical flow, and fusion-based mosaicking are presented from the first to the fourth column. The problematic parts of the panorama are denoted in blue, orange, and green rectangles from the first to the third column. The fusion-based mosaicking can correct them and combine advantages of the component methods to give high-quality panoramas.

of the pose graph as:

$$\begin{aligned}\mathbf{f} &= \sum_{i,j\in\mathcal{C}} \mathbf{e}_{ij}(\mathbf{x}_i,\mathbf{x}_j,\mathbf{z}_{i,j})^\top \boldsymbol{\Omega}_{ij}\mathbf{e}(\mathbf{x}_i,\mathbf{x}_j,\mathbf{z}_{i,j}) \\ &= \sum_{i,j\in\mathcal{C}^{flow}} \mathbf{e}_{ij}(\mathbf{x}_i,\mathbf{x}_j,\mathbf{z}^{flow}_{i,j})^\top \boldsymbol{\Omega}^{flow}_{ij}\mathbf{e}(\mathbf{x}_i,\mathbf{x}_j,\mathbf{z}^{flow}_{i,j}) \\ &+ \sum_{i,j\in\mathcal{C}^{SIFT}} \mathbf{e}_{ij}(\mathbf{x}_i,\mathbf{x}_j,\mathbf{z}^{SIFT}_{i,j})^\top \boldsymbol{\Omega}^{SIFT}_{ij}\mathbf{e}(\mathbf{x}_i,\mathbf{x}_j,\mathbf{z}^{SIFT}_{i,j}) \\ &+ \sum_{i,j\in\mathcal{C}^{reg}} \mathbf{e}_{ij}(\mathbf{x}_i,\mathbf{x}_j,\mathbf{z}^{reg}_{i,j})^\top \boldsymbol{\Omega}^{reg}_{ij}\mathbf{e}(\mathbf{x}_i,\mathbf{x}_j,\mathbf{z}^{reg}_{i,j})\end{aligned} \tag{10}$$

where $\mathcal{C}$ is set of all the edges including odometry edges ($j = i+1$) and loop closure edges ($j \neq i+1$), and the function $\mathbf{e}$ measures errors between the vertices and constraints by the edges. $\boldsymbol{\Omega}$ is the information matrix, *i.e.*, the inverse of the covariance matrix $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$. The covariance of an edge is decided based on the residual or number of correspondences of the two images. For the SIFT-based method, its covariance $\boldsymbol{\Sigma}^{SIFT}$ is inversely proportional to the number of established pairwise point correspondences. For the optical flow-based method, the covariance $\boldsymbol{\Sigma}^{flow}$ is inversely proportional to the number of RANSAC inliers as described in Section 3.2. For the registration-based method, the covariance $\boldsymbol{\Sigma}^{reg}$ is directly proportional to the photometric residual after optimisation. In this paper, we set the information matrix as follows: if the number of inliers or correspondences is N_{inl} for the SIFT-based or optical flow-based method, then the first four diagonal elements of the information matrix is set as $\boldsymbol{\Omega}_{(1,1)} = \boldsymbol{\Omega}_{(1,1)} = \cdots = \boldsymbol{\Omega}_{(4,4)} = 100 \times N_{inl}$, the last two diagonal elements are set as $\boldsymbol{\Omega}_{5,5} = \boldsymbol{\Omega}_{6,6} = N_{inl}$. If the residual for the direct registration-based method is e_{res}, then the first four diagonal elements are set as $\frac{100}{e_{res}}$, and the last two diagonal elements are $\frac{1}{e_{res}}$. If there is not enough inliers/correspondences, the residual is too large, or the output of the RANSAC algorithm is an identity matrix, then we set all elements of information matrix as zero

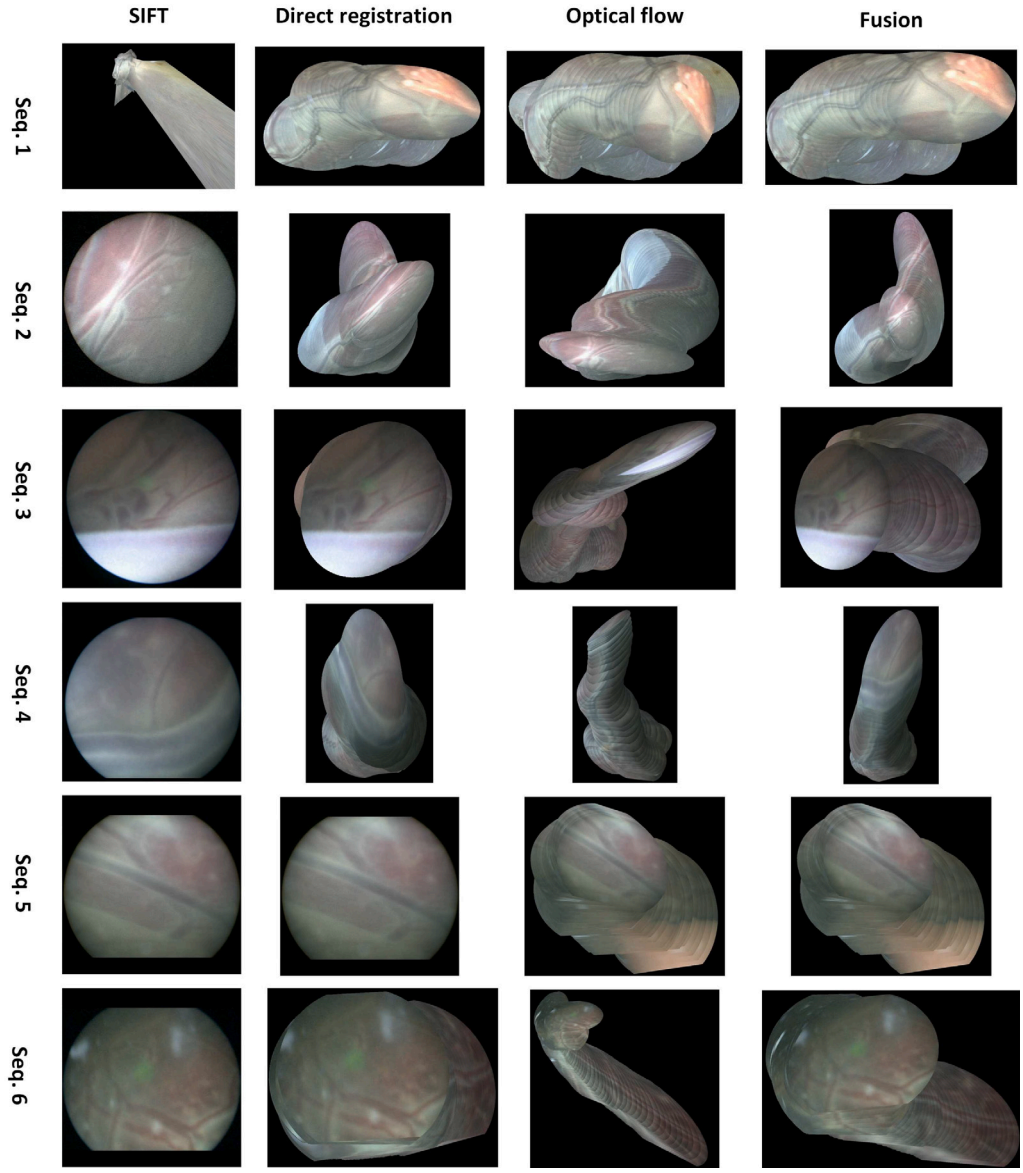


Fig. 6. Results on the fetoscopy dataset. Mosaicking results for six sequences are presented from the first to the last row. The SIFT, direct registration, optical flow, and fusion-based mosaicking are presented from the first to the fourth column. SIFT-based method fails to work on this dataset due to the texture-less background and difficulty to extract enough features. The fusion-based method fuses results of the direct registration-based and optical flow-based homography estimation, and can combine the advantages of both methods to generate better panoramas.

$\boldsymbol{\Omega} = \mathbf{0}$, which means we treat the estimation as a failure and do not consider the constraint provided by this edge. SIFT may also have loop closures to measure errors between non-consecutive vertices. Note that there may exist better information matrix configuration strategies, we left it to be explored in our future work. The error function $\mathbf{e}$ needs to be converted from the 3×3 affine matrix to a vector to compute and minimise the loss. Following our previous work (Li et al., 2021), the vectorisation is based on the Lie group theory, *i.e.*, from element on affine Lie group to its corresponding Lie algebra and then the vector space. And update of the state is wrapping from the vector space to Lie group. For a detailed elaboration please refer to Li et al. (2021), while in this paper, the key procedures are introduced briefly. We have $\mathbf{e}(\mathbf{x}_i, \mathbf{x}_j, \mathbf{z}_{i,j}) = \log(\mathbf{z}_{i,j}^{-1}\mathbf{x}_i^{-1}\mathbf{x}_j)^{\vee}$ using the logarithm map. Updating it with a small perturbation $\boldsymbol{\xi}$ in Lie algebra leads to $\mathbf{e}(\mathbf{x}\exp(\boldsymbol{\xi})) \simeq \mathbf{e}(\mathbf{x}) + \mathbf{J}\boldsymbol{\xi}$, where we take the first-order Taylor approximation and $\mathbf{J}$ is the Jacobian matrix from affine Lie group to vector space which can be computed by numerical method. The cost function on the updated state is

$$\begin{aligned}\mathbf{f}(\mathbf{x}\exp(\boldsymbol{\xi})) &\simeq \mathbf{f}(\mathbf{x}) + 2\underbrace{\sum_{i,j\in C}\mathbf{e}(\mathbf{x}_i,\mathbf{x}_j,\mathbf{z}_{ij})^{\top}\boldsymbol{\Omega}_{ij}\mathbf{J}_{ij}}_{\mathbf{b}^{\top}}\boldsymbol{\xi} \\ &+ \boldsymbol{\xi}^{\top}\underbrace{\sum_{i,j\in C}\mathbf{J}_{ij}^{\top}\boldsymbol{\Omega}_{ij}\mathbf{J}_{ij}}_{\mathbf{K}}\boldsymbol{\xi} = \mathbf{f}(\mathbf{x}) + 2\mathbf{b}^{\top}\boldsymbol{\xi} + \boldsymbol{\xi}^{\top}\mathbf{K}\boldsymbol{\xi}\end{aligned} \tag{11}$$

To make the update driven to the optimal value, we need to make the differential of the cost function equal to zero, *i.e.*, $\mathbf{f}(\mathbf{x}\exp(\boldsymbol{\xi})) - \mathbf{f}(\mathbf{x}) = 0$. The differential of equation (11) with respect to $\boldsymbol{\xi}$ is $\mathbf{K}\boldsymbol{\xi} + \mathbf{b} = \mathbf{0}$, which is linear. To improve the convergence, we adopt L–M algorithm here by incorporating a damping factor λ as $(\mathbf{K} + \lambda\mathbf{I})\boldsymbol{\xi}^{*} + \mathbf{b} = \mathbf{0}$. Then the state can be updated in this step as:

$$\mathbf{x}^{*} = \mathbf{x}\exp(\boldsymbol{\xi}^{*\wedge}) \tag{12}$$

The procedures from Eq. (11) to Eq. (12) iterate to update the state until convergence.

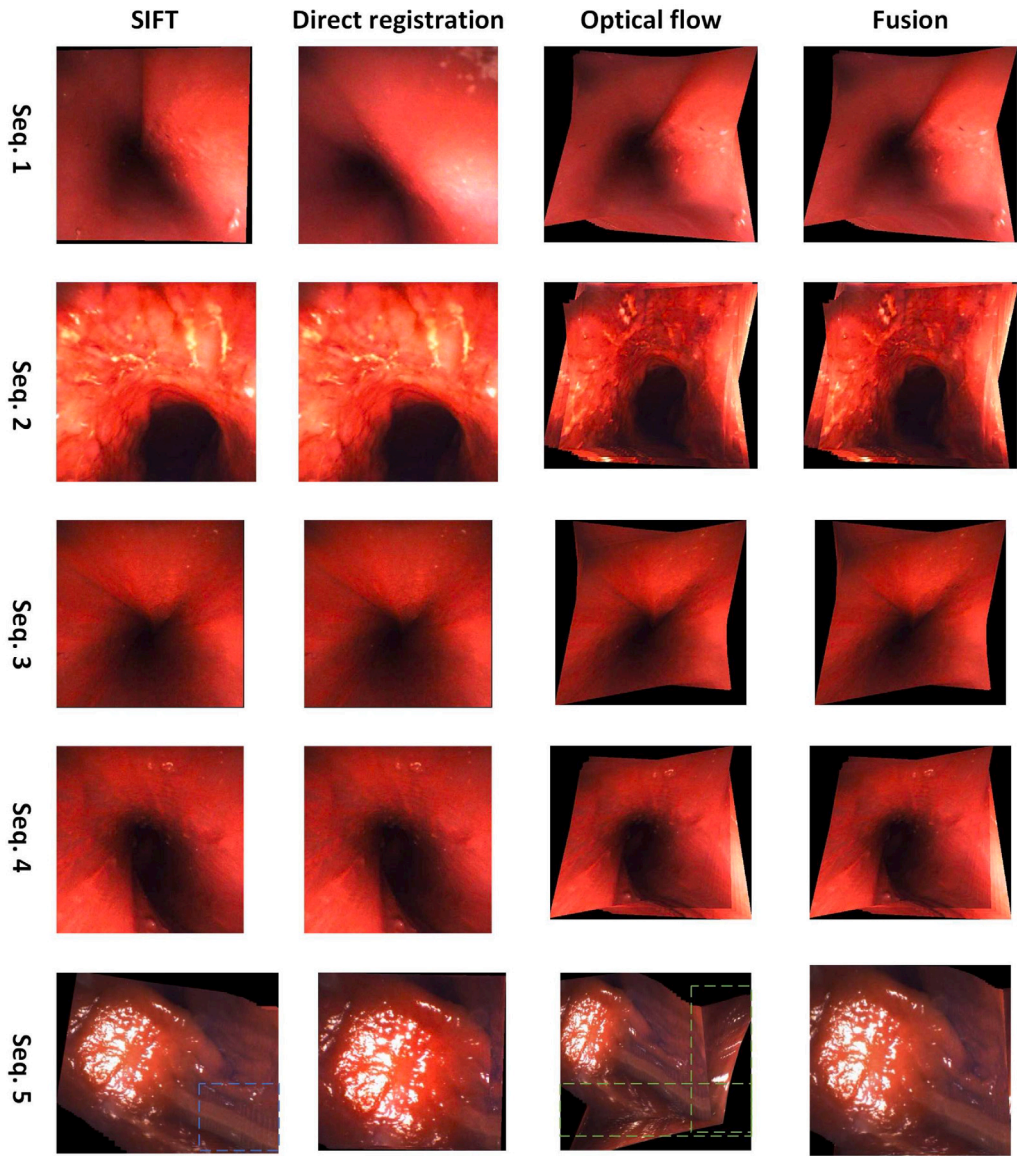


Fig. 7. Results on the human cadaver dataset. Mosaicking results for five sequences are presented from the first to the last row. The SIFT, direct registration, optical flow, and fusion-based mosaicking are presented from the first to the fourth column. From the first to the fourth sequence, only the optical flow works among the three component methods. And the result of fusion is same as that of optical-flow mosaicking. For the fifth sequence, the fusion-based method fuses the results of SIFT-based and optical flow-based homography estimation using the affine pose graph, to yield a more consistent panorama.

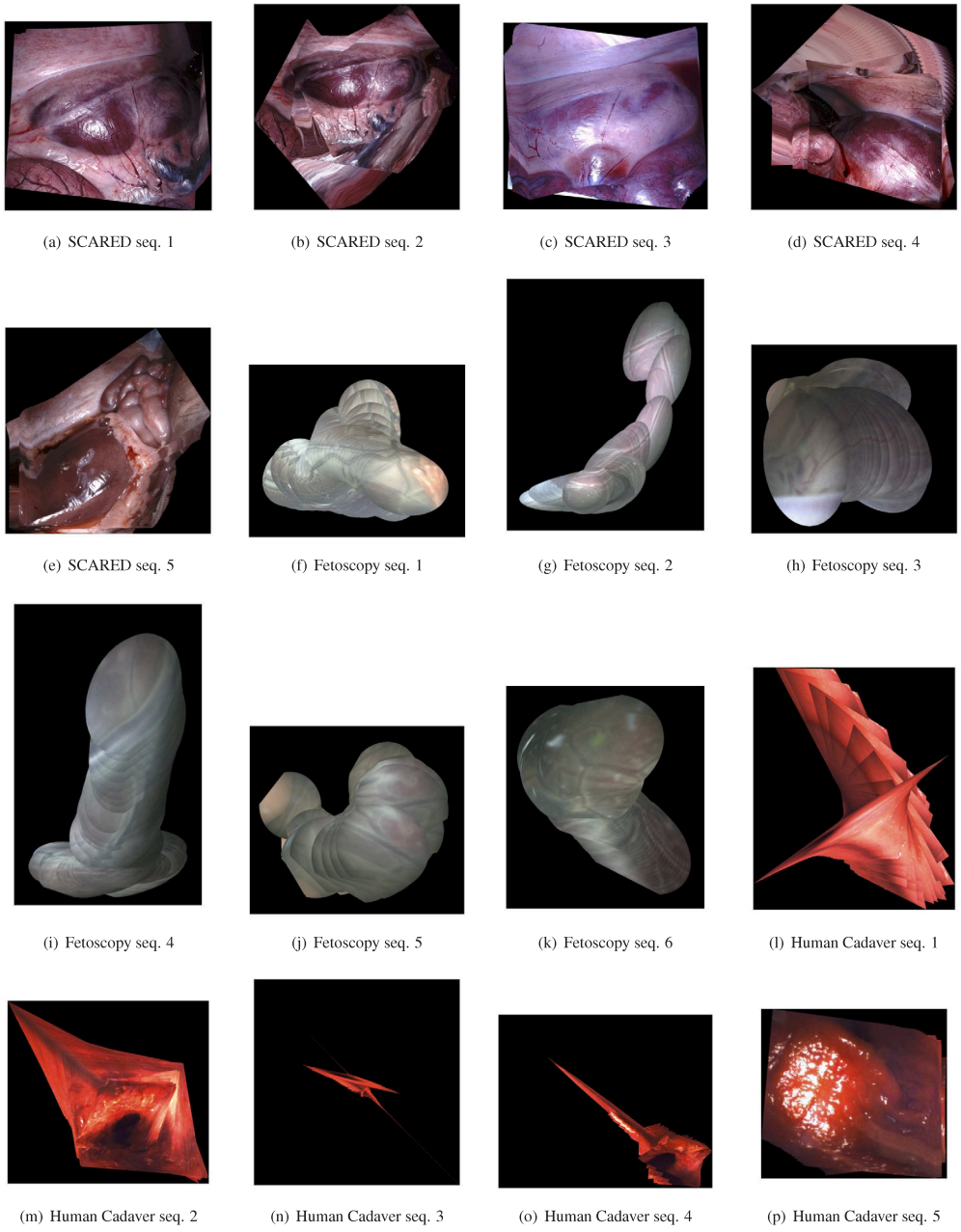
3.4. Panorama blending

Each image is attached with its own homography matrix with respect to its former image denoted as $\{I_i, \mathbf{H}_{i-1,i}\}, i = 1, 2, \ldots, n$, where n is the number of images and $\mathbf{H}_{0,1} = \mathbf{I}$ is the identity. To make a better visualisation, we set the transformation of the middle image ($\frac{n}{2}$ if n is an even number, $\frac{n+1}{2}$ otherwise) as the identity. Then every image can obtain its new transformation with respect to the middle image by matrix multiplication and inverse. For convenience, we use $\mathbf{H}_i$ to denote transformation of i-th image. First, we need to compute the resolution of the panorama. Every image is warped to its position using the affiliated homography matrix $\mathbf{H}_i$. The coordinates of four corners of the panorama can be obtained with minimal and maximal corners in two directions of all the wrapped images. In this way, we can create a blank mask that has the same size of the panorama. Then, for the first warped image, it can be fit into the mask directly. From the second image, only the mask's pixels that are still blank will be substituted with the pixels of the wrapped image. The proposed algorithm is summarised in Algorithm 1.

4. Experiments

In this section, experiments on various endoscopic datasets and comparison with state-of-the-art baselines will be presented. We test on three datasets: The first one is the Stereo Correspondence and Reconstruction of Endoscopic Data (SCARED) dataset (Allan et al., 2021). The utilised sequences are from its training data: Seq. 1 ← dataset1/keyframe1, Seq. 2 ← dataset1/keyframe2, Seq. 3 ← dataset2/keyframe1, Seq. 4 ← dataset3/keyframe3, Seq. 5 ← dataset4/keyframe4.

It was captured using stereo endoscopic cameras mounted on a da Vinci Xi surgical robot. This is a high quality, high resolution dataset with smooth camera motions. Nonetheless, illumination of different sequences on this dataset varies considerably. We use images from the left camera rather than the stereo in this paper. As we either need to treat the stereo pairs as sequences or blend the pairs first, which may cause new uncertainties. We note this dataset includes camera motion measurements provided by robot kinematics, however, these are not accurate enough for moosaicking (see results in Fig. 4) and



(a) SCARED seq. 1 (b) SCARED seq. 2 (c) SCARED seq. 3 (d) SCARED seq. 4

(e) SCARED seq. 5 (f) Fetoscopy seq. 1 (g) Fetoscopy seq. 2 (h) Fetoscopy seq. 3

(i) Fetoscopy seq. 4 (j) Fetoscopy seq. 5 (k) Fetoscopy seq. 6 (l) Human Cadaver seq. 1

(m) Human Cadaver seq. 2 (n) Human Cadaver seq. 3 (o) Human Cadaver seq. 4 (p) Human Cadaver seq. 5

Fig. 8. Mosaics generated by simple mean fusion of the SIFT-based, direct registration-based, and the optical flow-based estimation.

therefore we do not use this motion as a reference or groundtruth in this paper. The second is a fetoscopy placenta dataset (Bano et al., 2020) which has six sequences from different surgeries. Since this procedure is immersed in fluid, it does not contain specular reflections, but the scenes have very few discriminative textures and contain inconsistent motions due to floating particles crossing the field of view. By testing on this dataset, we want to verify generalisation of the proposed method in a significantly different environment and camera setting. The third dataset is a gastric endoscopy on a human cadaver using the Bellowscope gastroscope platform (Chandler et al., 2020). The environment is texture-less and with poor colour content, it has specular reflections, and has highly non-homogeneous illumination. Producing mosaics with this data represents the most extreme challenges for all studied algorithms. The numbers of frames of all the sequences of the three datasets are presented in Table 1. In this dataset, only small video sequences are tested. Due to the tubular shape of the anatomy, it does not make sense to build a single mosaic as the camera does a long trajectory through the digestive track, since it cannot be fully projected into a single plane without huge, non-intuitive distortions. No single algorithm works well in this setting. Instead, we do field of view expansion on localised portions of the anatomy where the endoscope is panning the scene.

We select three algorithms from the literature as comparison baselines, which correspond to mosaics as generated by the individual image registration approaches: feature-based (SIFT), direct pixel-based registration, and optical flow (FlowNet2.0). Each of them is a representative method in its own category (see Section 2). To further validate our approach, we also test our method when the covariance-weighted fusion is replaced with a naive simple average, and when loop closure is removed. In terms of quantitative analysis, we use the mosaicking metric described in Bano et al. (2019). This measures the structural similarity SSIM (Wang et al., 2004) between different overlapping

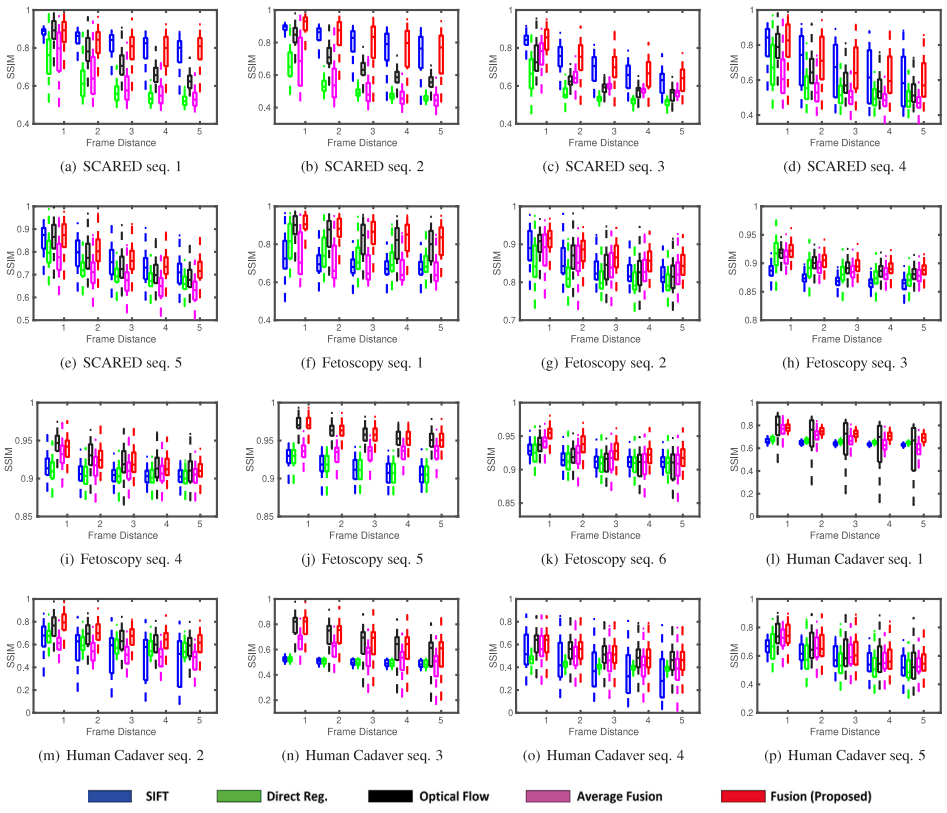


Fig. 9. SSIM between overlapping registered frames with distance between 1 (consecutive) and 5. Each boxplot shows SSIM results of all frame pairs in a video with specified distance. Lower values denote poorer methods.

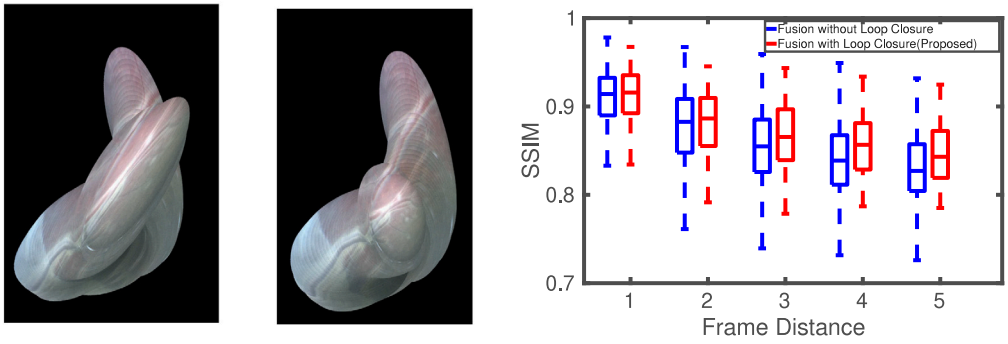


Fig. 10. A comparison of mosaicking generated by fusion with and without loop closure on sequence 2 of the fetoscopy dataset.

mosaicking frames across the entire sequences. We compare frames with increasing temporal distance from 1 (i.e. all consecutive frames) up to 5 frames apart.

4.1. Results

The qualitative mosaicking comparison against the baselines on the SCARED, fetoscopy, and human cadaver datasets are presented, respectively, in Figs. 5, 6, and 7. We present the results using the naive fusion scheme (simple average) separately in Fig. 8 for all datasets. In terms of quantitative results, we display the boxplots of SSIM distributions in Fig. 9 for frame distances between 1 and 5. We highlight that, while our proposed method establishes effective (long) loop closure constraints, these only occur in 3 out of the 6 fetoscopy datasets, and in none of the SCARED and human cadaver sequences due to the simple nature of the camera motions in these cases. A comparison of our method with and without loop closures for a sequence of the fetoscopy dataset is displayed in Fig. 10. Finally, the average SSIM results for all reported methods, across all datasets is summarised in Table 2.

To understand the contribution of the different baseline algorithms to our fusion we also provide their indicative weights for each dataset. On the SCARED dataset, the traces of the information matrix have orders of magnitude 10^5, 10^4, and 10^4 for the SIFT, direct registration, and optical flow respectively across the majority of image pairs. For the fetoscopy dataset, these values are 10^2, 10^5, and 10^6 respectively. For the human cadaver dataset 10^2, 10^2, and 10^4 for the first four sequences, and 10^4, 10^3, 10^4 for the fifth sequence. In this context, higher relative values mean that our fusion scheme is giving more importance to the respective method.

Algorithm 1: Specularity-aware Optical Flow-based Image Mosaicking

Input: Image sequence $\mathcal{I} = \{I_1, I_2, \dots, I_n\}$
Output: Panorama $\mathcal{P}$

1 **for** $i = 1,\ i < n$ **do**
2 $OpticalFlow_i = FlowNet2(I_i, I_{i+1})$ ▷ Optical flow prediction using the two images
3 **while** *not reaching maximal RANSAC step* **do**
4 $\mathbf{p}'_j = Random(I_i, OpticalFlow_i),\ j = 1, \dots, 4$ ▷ Randomly select four points
5 $\mathbf{p}_j = \mathbf{p}'_j + OpticalFlow_i^j$ ▷ Compute the corresponding points using equation (3)
6 $\hat{\mathbf{H}}_{i,i+1} = LS(\mathbf{p}_1, \mathbf{p}'_1, \dots, \mathbf{p}_4, \mathbf{p}'_4)$ ▷ Compute the homography matrix by solving equation (6)
7 $inliers = 0$ ▷ The initial number of inliers
8 **while** $k = 5, k \le N_{pairs}$ **do**
9 $\mathbf{d}_k = \|\mathbf{p}_k - \hat{\mathbf{H}}_{i,i+1}\mathbf{p}'_k\|$ ▷ Compute the residual error
10 **if** $\mathbf{d}_k < \varepsilon$ **then**
11 $inliers++$ ▷ Identity the remaining point is an inlier or outlier
12 **end**
13 **end**
14 **if** $inliers > \gamma$ **then**
15 $\mathbf{H}_{i,i+1}^{flow} = \hat{\mathbf{H}}_{i,i+1}$; break ▷ Stop iteration if number of inliers is large
16 **end**
17 **if** *reaching the last iteration* **then**
18 $\mathbf{H}_{i,i+1}^{flow} = Inliers_{max}(\{\hat{\mathbf{H}}_{i,i+1}\})$ ▷ Select the matrix that has most inliers as solution
19 **end**
20 **end**
21 $f_i^{SIFT} = SIFT(I_i),\ f_{i+1}^{SIFT} = SIFT(I_{i+1})$ ▷ Extract SIFT features from the images
22 $\mathbf{H}_{i,i+1}^{SIFT} = Aff_{ransac}(f_i^{SIFT}, f_{i+1}^{SIFT})$ ▷ Homography estimation using SIFT as line 3 - 20
23 $\mathbf{H}_{i,i+1}^{reg} = Reg(I_i, I_{i+1})$ ▷ Homography est. with direct registration as equation (9)
24 $Lib_{key} = BoG(f_{key}^{SIFT})$
25 **end**
▷ Construct the library of keyframe features using bag of words
26 $f_{key}^*, score = ImgRetr(f_i, Lib_{key})$ ▷ Compute the score the most likely frame by image retrieval
27 **if** *score is larger than the threshold* **then**
28 $\mathbf{H}_{i,key}^{LC} = Aff_{ransac}(f_i^{SIFT}, f_{key}^{SIFT})$
29 **end**
▷ Homography estimation of the loop closure $\mathcal{G} = PoseGraph(\{\mathbf{H}^{flow}, \boldsymbol{\Omega}^{flow}\}, \{\mathbf{H}^{SIFT}, \boldsymbol{\Omega}^{SIFT}\}, \{\mathbf{H}^{reg}, \boldsymbol{\Omega}^{reg}\})$ ▷ Pose graph construction
30 $\mathbf{x}_i = \prod_{k=1}^{i} \mathbf{H}_{k-1,k}^{v},\ v = \underset{m=\{flow,SIFT,reg\}}{\arg\min} \boldsymbol{\Sigma}_k^m$ ▷ Vertex is constructed by estimation with minimal covariance
31 $\mathbf{z}_{i,j}^{v} = \mathbf{H}_{i,j}^{v},\ v = \{flow, SIFT, reg\}$ ▷ Edges are constructed using the transformation estimation
32 $\mathbf{f} = \sum_{i,j \in \mathcal{C}} \mathbf{e}_{ij}(\mathbf{x}_i, \mathbf{x}_j, \mathbf{z}_{i,j})^\top \boldsymbol{\Omega}_{ij} \mathbf{e}(\mathbf{x}_i, \mathbf{x}_j, \mathbf{z}_{i,j})$ ▷ Cost function by vertices and edges as equation (10)
33 $\{\mathbf{H}_{i,i+1}\} = Opt(\mathcal{G})$ ▷ Affine pose graph optimisation
34 **while** *not converge* **do**
35 $\mathbf{J}_{ij} = \left.\frac{\partial \tilde{\mathbf{e}}_{ij}(\mathbf{x}\exp(\delta\boldsymbol{\xi}^\wedge))}{\partial \delta\boldsymbol{\xi}}\right|_{\delta\boldsymbol{\xi}=\mathbf{0}} \simeq \left.\frac{\tilde{\mathbf{e}}_{ij}(\mathbf{x}\exp(\delta\boldsymbol{\xi}^\wedge)) - \tilde{\mathbf{e}}(\mathbf{x})}{\delta\boldsymbol{\xi}}\right|_{\delta\boldsymbol{\xi}\to\mathbf{0}}$
36 **end**
▷ Compute the Jacobian using the numerical method
37 $\mathbf{f}_{ij}\left(\mathbf{x}\exp(\delta\boldsymbol{\xi}^\wedge)\right) \simeq (\tilde{\mathbf{e}}_{ij}(\mathbf{x}) + \mathbf{J}_{ij}\delta\boldsymbol{\xi})^\top \boldsymbol{\Omega}_{ij}(\tilde{\mathbf{e}}_{ij}(\mathbf{x}) + \mathbf{J}_{ij}\delta\boldsymbol{\xi})$ ▷ Add a small perturbation of f_{ij}
38 $\mathbf{f}(\mathbf{x}\exp(\boldsymbol{\xi})) = \sum_{i,j \in \mathcal{C}} \mathbf{f}_{ij} \simeq \mathbf{f}(\mathbf{x}) + 2\mathbf{b}^\top\boldsymbol{\xi} + \boldsymbol{\xi}^\top\mathbf{K}\boldsymbol{\xi}$ ▷ Approximation of the cost function as equation (11)
39 $(\mathbf{K} + \lambda\mathbf{I})\boldsymbol{\xi}^* + \mathbf{b} = \mathbf{0}$ ▷ Obtain the optimal update $\boldsymbol{\xi}^*$ using L-M algorithm
40 $\mathbf{x}^* = \mathbf{x}\exp(\boldsymbol{\xi}^{*\wedge})$ ▷ Update the estimation using equation (12)
41 **for** $i = 1,\ i \le n$ **do**
42 $\mathbf{H}_i = \begin{cases} \mathbf{H}_{m,m-1}^{-1} \cdots \mathbf{H}_{i,i+1}^{-1} & i < m \\ \mathbf{H}_{m,m+1} \cdots \mathbf{H}_{i-1,i} & i > m \end{cases}$ ▷ Compute the homography matrix w.r.t. the middle image
43 $I_i^w = Wrap(I_i, \mathbf{H}_i)$ ▷ Warp the image using its homography matrix
44 **end**
45 $\mathcal{P} = blending(I_1^w, I_2^w, \dots, I_n^w)$ ▷ Stitch the wrapped images to get the panorama

4.2. Discussion

In the SCARED dataset (Fig. 5), SIFT has generally better performance than direct registration and optical flow, which can be explained by the high resolution and rich textures that make it easy to extract keypoint features. For sequences 3 to 5, direct registration fails entirely to work and most images in the mosaic overlap completely (i.e. the registration outputs a 3×3 identity matrix). However, SIFT fails to find good features on sequences 4 and 5, resulting in bad quality results. From the fourth column of Fig. 5, we can see that our fusion approach

Table 2
A comparison of different methods and ablation study on the three datasets. Values in the table are average SSIM with frame distance from 1 to 5. Note that the ablation study of loop closure is tested using the first three sequences of the fetoscopy dataset as there are long loop closures in these sequences.

Dataset	Seq.	SIFT	Direct reg.	Optical flow	Average fusion	Fusion w/o LC	Fusion (proposed)
SCARED	1	0.822	0.607	0.741	0.631	N/A	0.841
	2	0.804	0.526	0.674	0.540	N/A	0.807
	3	0.709	0.597	0.620	0.633	N/A	0.718
	4	0.677	0.549	0.628	0.527	N/A	0.680
	5	0.770	0.712	0.756	0.690	N/A	0.780
Fetoscopy	1	0.697	0.746	0.834	0.691	0.856	0.857
	2	0.850	0.820	0.852	0.845	0.862	0.873
	3	0.872	0.888	0.896	0.891	0.894	0.901
	4	0.907	0.905	0.924	0.914	N/A	0.924
	5	0.915	0.915	0.960	0.912	N/A	0.960
	6	0.917	0.912	0.917	0.917	N/A	0.926
Hum. cad.	1	0.641	0.659	0.662	0.645	N/A	0.730
	2	0.537	0.594	0.656	0.521	N/A	0.681
	3	0.501	0.501	0.638	0.534	N/A	0.641
	4	0.407	0.416	0.520	0.453	N/A	0.520
	5	0.596	0.577	0.641	0.638	N/A	0.645

can remove errors of individual methods, relying on the ones with least covariance at any given point. In general, for this dataset, our fusion weights SIFT by an order of magnitude above the other two methods, which is consistent with the observed baseline performances. The quantitative results (Fig. 9) are also consistent with these results, showing our fusion having the best performance, closely followed by SIFT.

In the fetoscopy dataset (Fig. 6), the image resolution is not as high as that of the SCARED dataset. In addition, the environment is smooth and texture-less, which makes it difficult to extract keypoint features. Here, the SIFT-based mosaicking completely fails to work for all the six sequences (see the first column, the algorithm outputs a 3×3 identity matrix if there is not enough correspondences or inliers of the RANSAC method). The direct registration has a significant amount of drift, and the optical flow performs the best among the three baselines. From the last column of Fig. 6, we can see that the panorama generated by our proposed fusion performs the best. In general, our method provides lowest weights for SIFT estimations, and the highest to optical flow. The quantitative results in Fig. 9 also indicate that our fusion method produces in general higher SSIM scores, followed by optical flow.

In the human cadaver dataset (Fig. 7), the scene is mostly red and texture-less, which makes it very difficult to find correspondence or maximise similarity metrics. From sequence 1 to 4, both SIFT and direct registration fail to estimate the transformation between the images and cannot generate the panoramas. The optical flow-based mosaicking again performs the best out of the 3 baselines. In this dataset, our fusion also generally weights optical flow the highest. In fact, for the first four sequences, the fusion-based mosaicking results rely exclusively on the optical flow-based due to complete failure of other methods. For sequence 5, both SIFT and optical flow can generate a mosaic, but are not accurate in some regions (see the blue and green rectangles). The fusion-based method combines advantages of both the methods and gives a better panorama. For this challenging human cadaver dataset, the quality of the generated panorama is good with around 50 images using the proposed method.

The importance of weighting each method differently from frame-to-frame in our fusion approach is further validated by the fact that a simple average fusion works very poorly (see Fig. 8), which is further confirmed by the SSIM results in Fig. 9 and Table 2, where the average fusion is consistently close to the worst performers, since it is heavily contaminated by the worst of the three baselines at any given moment.

Finally, the effect of loop closure is the most significant on sequence 2 of the fetoscopy dataset (Fig. 10), where the camera performs a long trajectory before returning to the pre-visited area of the anatomy. Without loop closure, drift error is accumulated throughout the trajectory. When such a motion is not present (i.e. most of the other sequences), loop closure contributes little to the fusion performance.

All these experiments demonstrate the robustness of the proposed method and its generalisation across different datasets. Advantages of the proposed method over the state-of-the-art medical image mosaicking algorithms are validated through the comparison both with qualitative and quantitative results.

5. Conclusion

This paper presents a robust endoscopic image mosaicking framework based on fusion of multimodal estimation. One of the advantages of the proposed method is that it can work in different environments with no need to re-design the framework or finetune the parameters. Comparison with state-of-the-art baselines including SIFT-based, direct, end-to-end mosaicking shows that the proposed method is more robust to specular reflections or in feature-less environment. Moreover, the proposed framework is open to any other estimation method. It is rather straightforward to fit the new mosaicking methods into the proposed pose graph framework where only the evaluation of uncertainties of that method is needed. The limitations of the current framework include: It does not take the deformation into consideration; And it does not include the case that there may be outliers in the pose graph or inaccurate estimation of the uncertainties of the edges. In future work, we plan to solve these problems by developing an outlier-aware affine pose graph optimisation algorithm with deformation estimation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgements

This work was supported by the Wellcome/EPSRC Centre for Interventional and Surgical Sciences (WEISS) at UCL (203145Z/16/Z), EPSRC (EP/P027938/1), and H2020 FET (GA863146). Danail Stoyanov is supported by a Royal Academy of Engineering Chair in Emerging Technologies (CiET1819/2/36) and an EPSRC Early Career Research Fellowship (EP/P012841/1). Liang Li is supported by the National Natural Science Foundation of China (62203383) and (62088101).

References

Allan, M., Mcleod, J., Wang, C.C., Rosenthal, J.C., Fu, K.X., Zeffiro, T., Xia, W., Zhanshi, Z., Luo, H., Zhang, X., et al., 2021. Stereo correspondence and reconstruction of endoscopic data challenge. arXiv preprint arXiv:2101.01133.

Bano, S., Vasconcelos, F., Amo, M.T., Dwyer, G., Gruijthuijsen, C., Deprest, J., Ourselin, S., Vander Poorten, E., Vercauteren, T., Stoyanov, D., 2019. Deep sequential mosaicking of fetoscopic videos. In: Proc. Int. Conf. on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 311–319.

Bano, S., Vasconcelos, F., Shepherd, L.M., Vander Poorten, E., Vercauteren, T., Ourselin, S., David, A.L., Deprest, J., Stoyanov, D., 2020. Deep placental vessel segmentation for fetoscopic mosaicking. In: Proc. Int. Conf. on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 763–773.

Bartoli, A., 2008. Groupwise geometric and photometric direct image registration. IEEE Trans. Pattern Anal. Mach. Intell. 30 (12), 2098–2108.

Baum, Z.M., Hu, Y., Barratt, D.C., 2021. Real-time multimodal image registration with partial intraoperative point-set data. Med. Image Anal. 102231.

Bay, H., Tuytelaars, T., Van Gool, L., 2006. Surf: Speeded up robust features. In: Proc. European Conf. Computer Vision. Springer, pp. 404–417.

Bergen, T., Ruthotto, S., Munzenmayer, C., Rupp, S., Paulus, D., Winter, C., 2009. Feature-based real-time endoscopic mosaicking. In: Proc. Int. Symp. Image and Signal Processing and Analysis. IEEE, pp. 695–700.

Can, A., Stewart, C.V., Roysam, B., Tanenbaum, H.L., 2002a. A feature-based, robust, hierarchical algorithm for registering pairs of images of the curved human retina. IEEE Trans. Pattern Anal. Mach. Intell. 24 (3), 347–364.

Can, A., Stewart, C.V., Roysam, B., Tanenbaum, H.L., 2002b. A feature-based technique for joint, linear estimation of high-order image-to-mosaic transformations: mosaicing the curved human retina. IEEE Trans. Pattern Anal. Mach. Intell. 24 (3), 412–419.

Capek, M., Krekule, I., 1999. Alignment of adjacent picture frames captured by a CLSM. IEEE Trans. Inf. Technol. Biomed. 3 (2), 119–124.

Chandler, J.H., Chauhan, M., Caló, S., Aruparayil, N., Garbin, N., Campisano, F., Obstein, K.L., Valdastri, P., 2020. Tu1964 usability of a novel disposable endoscope for gastric cancer screening in low-resource settings: results from rural India. Gastroenterology 158 (6), S–1235.

Chaudhari, K., Garg, D., Kotecha, K., 2017. An enhanced approach in image mosaicing using ORB method with alpha blending technique. Int. J. Adv. Res. Comput. Sci. 8 (5).

Chon, J., Fuse, T., Shimizu, E., Shibasaki, R., 2007. Three-dimensional image mosaicking using multiple projection planes for 3-D visualization of roadside standing buildings. IEEE Trans. Syst. Man Cybern. B 37 (4), 771–783.

Chum, O., Matas, J., 2012. Homography estimation from correspondences of local elliptical features. In: Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012). IEEE, pp. 3236–3239.

Daga, P., Chadebecq, F., Shakir, D.I., Herrera, L.C.G.-P., Tella, M., Dwyer, G., David, A.L., Deprest, J., Stoyanov, D., Vercauteren, T., et al., 2016. Real-time mosaicing of fetoscopic videos using SIFT. In: Medical Imaging 2016: Image-Guided Procedures, Robotic Interventions, and Modeling. 9786, International Society for Optics and Photonics, p. 97861R.

Dosovitskiy, A., Fischer, P., Ilg, E., Hausser, P., Hazirbas, C., Golkov, V., Van Der Smagt, P., Cremers, D., Brox, T., 2015. Flownet: Learning optical flow with convolutional networks. In: Proc. IEEE Int. Conf. Computer Vis.. pp. 2758–2766.

Garbin, N., Mamunes, A.P., Sohn, D., Hawkins, R.W., Valdastri, P., Obstein, K.L., 2019. Evaluation of a novel low-cost disposable endoscope for visual assessment of the esophagus and stomach in an ex-vivo phantom model. Endosc. Int. Open 7 (09), E1175–E1183.

Garbin, N., Wang, L., Chandler, J.H., Obstein, K.L., Simaan, N., Valdastri, P., 2018. Dual-continuum design approach for intuitive and low-cost upper gastrointestinal endoscopy. IEEE Trans. Biomed. Eng. 66 (7), 1963–1974.

Geiger, A., Lenz, P., Urtasun, R., 2012. Are we ready for autonomous driving? the kitti vision benchmark suite. In: Proc. IEEE Conf. Computer Vis. Pattern Recogn.. IEEE, pp. 3354–3361.

Gong, L., Zheng, J., Ping, Z., Wang, Y., Wang, S., Zuo, S., 2021. Robust mosaicing of endomicroscopic videos via context-weighted correlation ratio. IEEE Trans. Biomed. Eng. 68 (2), 579–591.

Gupta, S., Chakarvarti, S., Zaheeruddin, 2016. Medical image registration based on fuzzy c-means clustering segmentation approach using SURF. Int. J. Biomed. Eng. Technol. 20 (1), 33–50.

Huang, T.S., Netravali, A.N., 2002. Motion and structure from feature correspondences: A review. Adv. Image Process. Underst.: A Festschrift for Thomas S Huang 331–347.

Ilg, E., Mayer, N., Saikia, T., Keuper, M., Dosovitskiy, A., Brox, T., 2017. Flownet 2.0: Evolution of optical flow estimation with deep networks. In: Proc. IEEE Conf. Computer Vision Pattern Recogn.. pp. 2462–2470.

Jalili, J., Hejazi, S.M., Riazi-Esfahani, M., Eliasi, A., Ebrahimi, M., Seydi, M., Fard, M.A., Ahmadian, A., 2020. Retinal image mosaicking using scale-invariant feature transformation feature descriptors and voronoi diagram. J. Med. Imaging 7 (4), 044001.

Konen, W., Tombrock, S., Scholz, M., 2007. Robust registration procedures for endoscopic imaging. Med. Image Anal. 11 (6), 526–539.

Kutarnia, J., Pedersen, P., 2015. A Markov random field approach to group-wise registration/mosaicing with application to ultrasound. Med. Image Anal. 24 (1), 106–124.

Lee, S.-C., Bajcsy, P., 2005. Feature based registration of fluorescent LSCM imagery using region centroids. In: Proc. Medical Imaging: Image Processing. 5747, International Society for Optics and Photonics, pp. 170–181.

Levin, A., Zomet, A., Peleg, S., Weiss, Y., 2004. Seamless image stitching in the gradient domain. In: Proc. European Conf. Computer Vision. Springer, pp. 377–389.

Li, L., Bano, S., Deprest, J., David, A.L., Stoyanov, D., Vasconcelos, F., 2021. Globally optimal fetoscopic mosaicking based on pose graph optimisation with affine constraints. IEEE Robot. Autom. Lett. 6 (4), 7831–7838.

Li, D., He, Q., Liu, C., Yu, H., 2017. Medical image stitching using parallel sift detection and transformation fitting by particle swarm optimization. J. Med. Imag. Health Inform. 7 (6), 1139–1148.

Li, Y., Wang, Y., Huang, W., Zhang, Z., 2008. Automatic image stitching using SIFT. In: Proc. Int. Conf. Audio, Language Image Processing. IEEE, pp. 568–571.

Loewke, K.E., Camarillo, D.B., Piyawattanametha, W., Mandella, M.J., Contag, C.H., Thrun, S., Salisbury, J.K., 2010. In vivo micro-image mosaicing. IEEE Trans. Biomed. Eng. 58 (1), 159–171.

Loewke, N.O., Qiu, Z., Mandella, M.J., Ertsey, R., Loewke, A., Gunaydin, L.A., Rosenthal, E.L., Contag, C.H., Solgaard, O., 2020. Software-based phase control, video-rate imaging, and real-time mosaicing with a lissajous-scanned confocal microscope. IEEE Trans. Med. Imaging 39 (4), 1127–1137.

Lowe, D.G., 1999. Object recognition from local scale-invariant features. In: Proc. IEEE Int. Conf. Computer Vision. 2, Ieee, pp. 1150–1157.

Mayer, N., Ilg, E., Hausser, P., Fischer, P., Cremers, D., Dosovitskiy, A., Brox, T., 2016. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In: Proc. IEEE Conf. Computer Vis. Pattern Recogn.. pp. 4040–4048.

Milgram, D.L., 1975. Computer methods for creating photomosaics. IEEE Trans. Comput. 100 (11), 1113–1119.

Miranda-Luna, R., Daul, C., Blondel, W.C., Hernandez-Mier, Y., Wolf, D., Guillemin, F., 2008. Mosaicing of bladder endoscopic image sequences: Distortion calibration and registration algorithm. IEEE Trans. Biomed. Eng. 55 (2), 541–553.

Nguyen, T., Chen, S.W., Shivakumar, S.S., Taylor, C.J., Kumar, V., 2018. Unsupervised deep homography: A fast and robust homography estimation model. IEEE Robot. Autom. Lett. 3 (3), 2346–2353.

Ogien, J., Daures, A., Cazalas, M., Levecq, O., Dubois, A., 2020. Video-mosaicking of human skin in vivo using handheld line-field confocal optical coherence tomography. In: Photonics in Dermatology and Plastic Surgery 2020. 11211, International Society for Optics and Photonics, 1121114.

Okumura, K.-i., Raut, S., Gu, Q., Aoyama, T., Takaki, T., Ishii, I., 2013. Real-time feature-based video mosaicing at 500 fps. In: Proc. IEEE/RSJ Int. Conf. on Intell. Robots and Systems. IEEE, pp. 2665–2670.

Oliveira, M., Sappa, A.D., Santos, V., 2015. A probabilistic approach for color correction in image mosaicking applications. IEEE Trans. Image Process. 24 (2), 508–523.

Peter, L., Tella-Amo, M., Shakir, D.I., Attilakos, G., Wimalasundera, R., Deprest, J., Ourselin, S., Vercauteren, T., 2018. Retrieval and registration of long-range overlapping frames for scalable mosaicking of in vivo fetoscopy. Int. J. Comput. Assist. Radiol. Surg. 13 (5), 713–720.

Reeff, M., Gerhard, F., Cattin, P., Gábor, S., 2006. Mosaicing of endoscopic placenta images. INFORMATIK 2006–Informatik FÜR Menschen, Band 1.

Richa, R., Linhares, R., Comunello, E., Von Wangenheim, A., Schnitzler, J.-Y., Wassmer, B., Guillemot, C., Thuret, G., Gain, P., Hager, G., et al., 2014. Fundus image mosaicking for information augmentation in computer-assisted slit-lamp imaging. IEEE Trans. Med. Imaging 33 (6), 1304–1312.

Richa, R., Vágvölgyi, B., Balicki, M., Hager, G., Taylor, R.H., 2012. Hybrid tracking and mosaicking for information augmentation in retinal surgery. In: Proc. Int. Conf. Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 397–404.

Rong, W., Chen, H., Liu, J., Xu, Y., Haeusler, R., 2009. Mosaicing of microscope images based on SURF. In: Proc. Int. Conf. Image Vision Computing New Zealand. IEEE, pp. 271–275.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. In: Proc. Int. Conf. on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 234–241.

Rosa, B., Dahroug, B., Tamadazte, B., Rabenorosoa, K., Rougeot, P., Andreff, N., Renaud, P., 2018. Online robust endomicroscopy video mosaicking using robot prior. IEEE Robot. Autom. Lett. 3 (4), 4163–4170.

Seshamani, S., Lau, W., Hager, G., 2006. Real-time endoscopic mosaicking. In: Proc. Int. Conf. Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 355–363.

Seshamani, S., Smith, M.D., Corso, J.J., Filipovich, M.O., Natarajan, A., Hager, G.D., 2009. Direct global adjustment methods for endoscopic mosaicking. In: Medical Imaging 2009: Visualization, Image-Guided Procedures, and Modeling. 7261, International Society for Optics and Photonics, p. 72611D.

Soper, T.D., Porter, M.P., Seibel, E.J., 2012. Surface mosaics of the bladder reconstructed from endoscopic video for automated surveillance. IEEE Trans. Biomed. Eng. 59 (6), 1670–1680.

Sotiras, A., Davatzikos, C., Paragios, N., 2013. Deformable medical image registration: A survey. IEEE Trans. Med. Imaging 32 (7), 1153–1190.

Tian, F., Shi, P., 2014. Image mosaic using ORB descriptor and improved blending algorithm. In: 2014 7th International Congress on Image and Signal Processing. IEEE, pp. 693–698.

Viergever, M.A., Maintz, J.A., Klein, S., Murphy, K., Staring, M., Pluim, J.P., 2016. A survey of medical image registration–under review. Med. Image Anal. 33, 140–144.

Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P., 2004. Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. 13 (4), 600–612.

Wang, X., Sun, J., Peng, H.-Y., 2012. Efficient panorama mosaicing based on enhanced-FAST and graph cuts. In: Recent Advances in Computer Science and Information Engineering. Springer, pp. 757–762.

Wu, C.H., Su, M.Y., 2017. Specular highlight detection from endoscopic images for shape reconstruction. In: Applied Mechanics and Materials. 870, Trans Tech Publ, pp. 357–362.

Xu, Q., Chen, J., Luo, L., Gong, W., Wang, Y., 2020. UAV image mosaicking based on multiregion guided local projection deformation. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 13, 3844–3855.

Yu, D., Yang, F., Yang, C., Leng, C., Cao, J., Wang, Y., Tian, J., 2015. Fast rotation-free feature-based image registration using improved N-SIFT and GMM-based parallel optimization. IEEE Trans. Biomed. Eng. 63 (8), 1653–1664.

Zhang, J., Huang, Y., Song, Y., Jiang, Y., Zhang, L., Zhang, Y., 2019. Convolutional neural network-based registration for mosaicing of microscopic images. J. Electron. Imaging 28 (4), 043006.

Zhou, H., Jayender, J., 2021. Real-time nonrigid mosaicking of laparoscopy images. IEEE Trans. Med. Imaging 40 (6), 1726–1736.