Journal of Medical Imaging

MedicalImaging.SPIEDigitalLibrary.org

Pruning strategies for efficient online globally consistent mosaicking in fetoscopy

Marcel Tella-Amo
Loïc Peter
Dzhoshkun I. Shakir
Jan Deprest
Danail Stoyanov
Tom Vercauteren
Sebastien Ourselin

SPIE.

Marcel Tella-Amo, Loïc Peter, Dzhoshkun I. Shakir, Jan Deprest, Danail Stoyanov, Tom Vercauteren, Sebastien Ourselin, "Pruning strategies for efficient online globally consistent mosaicking in fetoscopy," *J. Med. Imag.* **6**(3), 035001 (2019), doi: 10.1117/1.JMI.6.3.035001.

Journal of Medical Imaging 6(3), 035001 (Jul–Sep 2019)

Pruning strategies for efficient online globally consistent mosaicking in fetoscopy

Marcel Tella-Amo,[a,*] Loïc Peter,[a] Dzhoshkun I. Shakir,[b] Jan Deprest,[c] Danail Stoyanov,[a] Tom Vercauteren,[b] and Sebastien Ourselin[b]
[a]UCL, WEISS, London, United Kingdom
[b]King's College London, School of Biomedical Engineering and Imaging Sciences, London, United Kingdom
[c]KU Leuven, Department of Development and Regeneration, Leuven, Belgium

Abstract. Twin-to-twin transfusion syndrome is a condition in which identical twins share a certain pattern of vascular connections in the placenta. This leads to an imbalance in the blood flow that, if not treated, may result in a fatal outcome for both twins. To treat this condition, a surgeon explores the placenta with a fetoscope to find and photocoagulate all intertwin vascular connections. However, the reduced field of view of the fetoscope complicates their localization and general overview. A much more effective exploration could be achieved with an online mosaic created at exploration time. Currently, accurate, globally consistent algorithms such as bundle adjustment cannot be used due to their offline nature, while online algorithms lack sufficient accuracy. We introduce two pruning strategies facilitating the use of bundle adjustment in a sequential fashion: (1) a technique that efficiently exploits the potential of using an electromagnetic tracking system to avoid unnecessary matching attempts between spatially inconsistent image pairs, and (2) an aggregated representation of images, which we refer to as superframes, that allows decreasing the computational complexity of a globally consistent approach. Quantitative and qualitative results on synthetic and phantom-based datasets demonstrate a better trade-off between efficiency and accuracy. [DOI: 10.1117/1.JMI.6.3.035001]

Keywords: mosaicking; fetoscopy; electromagnetic; twin-to-twin transfusion syndrome; drift-free; efficient.

Paper 19038R received Feb. 14, 2019; accepted for publication Jul. 9, 2019; published online Aug. 7, 2019.

1 Introduction

Twin-to-twin transfusion syndrome (TTTS) is a fetal condition that affects monochronic diamniotic pregnancies in which the presence of a certain pattern of intertwin vascular connections, known as anastomoses, results in an imbalance of the blood flow between twins. If this condition remains untreated, the outcome is generally fatal for both twins.[1] Minimally invasive surgery is today the standard of care. The surgeon explores the placenta using a fetoscope with a small field of view to find and photocoagulate all anastomoses. Despite the wide literature in endoscopic scenarios,[2,3] the challenges found in fetoscopic imagery complicate the application of available techniques. To illustrate the complexity of the fetoscopic site, Fig. 1 shows three examples of TTTS procedure where the reduced field of view can be observed. The surgeon must remember the explored areas to navigate through the placental surface building a mental map of the placenta. This is an extremely challenging task even for the best-trained surgeons due to this limited field of view, lack of texture, and turbidity, which hinders the visualization and also precludes any assistance from others. As a solution, mosaicking has been suggested to increase the field of view by stitching the images together in a common space, forming a map of the area.

The online creation of such map would allow for a more effective exploration[4,5] and localization of the anastomoses. Online approaches for mosaicking[6–8] usually rely on approximations that either summarize past information or do not make use of all available information. In some applications, a simple pairwise estimation,[5,9] where subsequent images are registered, can be appropriate for real-time operation. However, these approaches accumulate drift.

To mitigate the drift, a globally consistent method might be used. Bundle adjustment[10] is considered to be the offline reference method in globally consistent mosaicking due to its accuracy, which comes from the effective use of all the available information in a probabilistic way. When the number of frames increases, however, the limitations in computational complexity become more evident, and online operation becomes prohibitive.

In Ref. 6, the authors proposed to use an electromagnetic tracker (EMT) to guide the estimation and mitigate the drift. However, due to their computational complexity or suboptimality, the proposed strategies were not suitable to obtain accurate online mosaics with clinically acceptable update times.

Some efforts have been made to reduce the computational time of globally consistent approaches. For example, Schroeder et al.[11] proposed closed-form initial estimates to accelerate its convergence. Steedly et al.[12] used the concept of keyframes in an offline setting to denote a set of the most important frames; an all-to-all strategy analogous to bundle adjustment is used with the keyframes whereas a pairwise web connects the rest of frames enforcing only local consistency between them. This permits to reach mosaics of the order of thousands of frames. Despite the computational advantage of this approach, the reduction in the number of connections between nonkeyframe images leads to a decrease in robustness given that not all the information available has been used. The use of the redundancy provided by nonkeyframe images becomes essential

*Address all correspondence to Marcel Tella-Amo, E-mail: marcel.tella.14@ucl.ac.uk

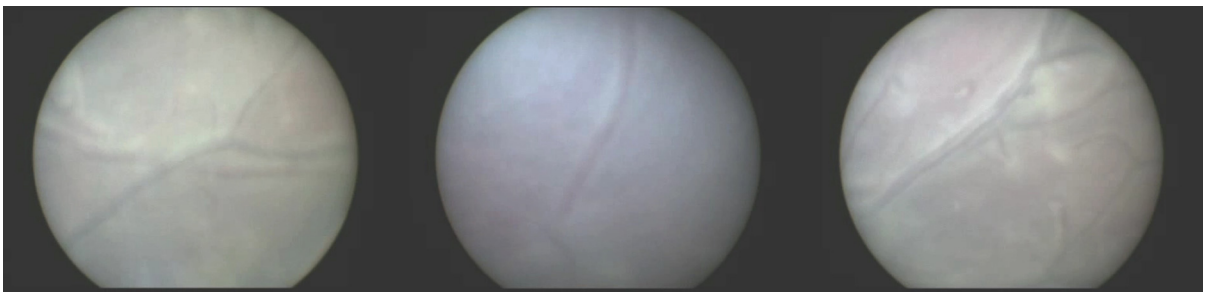

Fig. 1 Three examples of images taken in clinical conditions from a TTTS surgery.

when dealing with fetoscopic images due to their relatively poor quality.

An alternative to Ref. 12 consists of matching exclusively the images that are highly likely to share visual corresponding points, avoiding incoherent attempts. A proposed technique for overlap detection in monocular systems is to use a bag of words[13] (BoW), which creates a dictionary of visual words and assigns a signature per image. This signature summarizes the number of occurrences of the visual words in a histogram. In the case that the scene is revisited, the signature of the current image is very similar to the signature of the images initially acquired. However, this strategy works in an offline fashion while we aim for a sequential approach. This is the case for the work of Garcia-Fidalgo et al.[14] where an incremental BoW approach is used. However, pose information provided by an EMT system is much stronger than visual information in the placenta given that some uniform textureless areas of the placenta can lead to challenging scenarios. Therefore, we propose to use the EMT system to infer the topology of the cameras imaging the scene.

The organization of this paper is as follows: In Sec. 2, we review the background in mosaicking using the EMT system, to further detail our methodological contributions. The presentation of our experimental suite is done in Sec. 3, demonstrating a decrease in the computational complexity of bundle adjustment while maintaining similar accuracy. Then, we comment on the results in Sec. 4 and conclude the study in Sec. 5.

2 Methods

In this section, we present the preliminary background to then detail our algorithms. We start by introducing the scenario and stating the assumptions made to simplify the problem throughout the paper.

- We consider a placenta to be a plane. In Ref. 15, the authors demonstrated that this assumption greatly simplifies the problem.
- We assume the placenta to be static.
- Despite the image is obtained using a fetoscope, we use a pinhole camera model.[13,15]
- The EMT field is not perfect. Yet, we assume we are working on the center of the field, and that therefore, that the distortions due to inhomogeneities in the electromagnetic field are neglibible.[6]
- We assume[6] the each EMT measurement $\mathbf{z}_k$ to be centered on the true measurement $\mathbf{x}_k$, with a multivariate Gaussian noise of covariance matrix Σ_{EMT}.

We now introduce some concepts that are essential for the rest of the paper: how the visual aligned is performed in a pairwise and globally consistent manner, the link between poses and their imaged scenes over a planar surface, and how the EMT system can be introduced into the mosaicking pipeline.

2.1 *Background in Mosaicking*

Given a sequence of K images $\mathcal{I} = \{\mathbf{I}_k\}_{k=1}^{K}$, the goal of mosaicking is to find a two-dimensional representation of the scene or mosaic $\mathbf{M}:\Omega_M \rightarrow \mathbb{R}^3$ (RGB) where $\Omega_M \subset \mathbb{R}^2$ is denoted the mosaic space. Provided that the camera observes a planar scene, a homography $\mathbf{H}$ exists between corresponding points in two views $\mathbf{p} = [\,p_x \quad p_y\,]^T$, $\mathbf{p}' = [\,p'_x \quad p'_y\,]^T$, which lie in their respective image spaces Ω_p, $\Omega_{p'} \subset \mathbb{R}^2$. This homography can be parametrized as a 3×3 matrix such that

$$p'_x = \frac{H_{1,1}p_x, +H_{1,2}p_y + H_{1,3}}{H_{3,1}p_x, +H_{3,2}p_y, +H_{3,3}},$$
$$p'_y = \frac{H_{2,1}p_x + H_{2,2}p_y + H_{2,3}}{H_{3,1}p_x + H_{3,2}p_y + H_{3,3}}. \tag{1}$$

Then, Eq. (1) can be written as $\tilde{\mathbf{p}}' \propto \mathbf{H}\tilde{\mathbf{p}}$ where $\mathbf{H}$ is defined up to scale and the tilde indicates that the points are expressed in homogeneous coordinates.[16]

If the same planar scene is observed from both views, a homography $\mathbf{H}$ can be directly inferred from image matching information.[17] In this work, we use a landmark-based registration approach[18–20] since these approaches usually allow for sparse feature detection which can be faster than using the information in the whole image.

Pairwise mosaicking (PM) relies on estimations of pairwise homographies $\mathbf{H}_{k+1,k}$ which project any point from the space Ω_k of image $\mathbf{I}_k$ onto the space Ω_{k+1} of image $\mathbf{I}_{k+1}$. Pairwise homographies are then used to compose homographies from a fixed reference, which without loss of generality, can be placed on the first frame as

$$\mathbf{H}_k = \mathbf{H}_{k,1} = \prod_{j=1}^{k-1} \mathbf{H}_{j+1,j}, \tag{2}$$

where the product operator denotes the left matrix multiplication, corresponding to the composition of homographies, e.g., $\mathbf{H}_3 = \mathbf{H}_{3,1} = \mathbf{H}_{3,2}\mathbf{H}_{2,1}$. Once the homographies relating every frame to the reference are computed, every pixel in every image is projected onto the mosaic space. To further clarify the nomenclature, we name absolute homographies $\mathbf{H}_k$ with a single subindex meaning that they project any point from the reference Ω_M

to the image frame space Ω_k, whereas pairwise homographies $\mathbf{H}_{k+1,k}$ are defined as mapping points from Ω_k to Ω_{k+1}. When performing PM, any residual error in the estimation of pairwise homographies is accumulated through the chain in Eq. (2), resulting in a wrong placement of the images in the mosaic that degenerates in an uncontrolled way over time.

Globally consistent approaches such as bundle adjustment[10] take into account the relationship between all pairs of images in the sequence to create globally consistent mosaics.[21] However, they are generally not fast enough for online operation; the main reasons are the acquisition of the correspondences between all images and the run-time of the nonlinear optimization procedure.

Let us define $\mathcal{L}$ as the set of all pairs of image indices in which correspondences have been successfully acquired. We aim to obtain a set of estimated homographies $\hat{\mathbf{H}}_1, \ldots, \hat{\mathbf{H}}_{K-1}$, where the hat denotes that they are an estimate, that minimizes the reprojection errors of the matching points in all images in the sequence, namely:

$$\hat{\mathbf{H}}_1, \ldots, \hat{\mathbf{H}}_{K-1} = \underset{\mathbf{H}_1, \ldots, \mathbf{H}_{K-1}}{\arg\min} \sum_{\{l,m\} \in \mathcal{L}} \sum_{i=1}^{N_{l,m}} \frac{1}{\sigma_v^2} \| \mathbf{p}_l^i - f(\mathbf{H}_l \mathbf{H}_m^{-1} \tilde{\mathbf{p}}_m^i) \|_2^2. \tag{3}$$

$N_{l,m}$ is the number of matches found in the pair $\{l, m\}$, $f(\cdot)$ is the conversion from homogeneous to Cartesian coordinates so that $\mathbf{p} = f(\tilde{\mathbf{p}})$, and σ_v^2 is the variance associated with the location of a feature once propagated to the space of its matched pair. This variance is associated with the fact that we model the error between a fixed point in an image, and its corresponding pair propagated from the other image as a Gaussian random variable. Note that, since one frame is defined as the reference (here chosen as the first frame, without loss of generality), we only require to estimate $K - 1$ homographies, the remaining one being set to identity.

Let $\mathcal{X} = \{\mathbf{x}_k\}_{k=1}^K$ be the set of corresponding true camera poses. In addition, let $\mathcal{Z} = \{\mathbf{z}_k\}_{k=1}^K$ be the respective EMT measurements of these poses, where we assume $\mathbf{z}_k$ to be a noisy instance of the true camera $\mathbf{x}_k$. We parametrize each camera pose $\mathbf{x}_k = [\mathbf{r}_k, \mathbf{t}_k]^T$ as a rotation $\mathbf{r}_k$ and translation $\mathbf{t}_k$. The vector $\mathbf{z}_k$ is parametrized in the same way. The three-parameter rotation vector $\mathbf{r}_k$ is extracted from the skew symmetric matrices $[\mathbf{r}_k]_x \in \mathfrak{so}(3)$, which can be converted into the rotation matrix $\mathbf{R}_k = \exp([\mathbf{r}_k]_x) \in SO(3)$.[22] $SO(3)$ refers to the special orthogonal group of matrices of dimension 3×3 with determinant 1, and $\mathfrak{so}(3)$ is its corresponding lie algebra. With a rotation matrix $\mathbf{R}_k$ and a translation $\mathbf{t}_k$, we can compose the rigid transformation $\mathbf{T}_k \in SE(3)$, where $SE(3)$ is the special Euclidean group corresponding to rigid transformations in three-dimensional space

$$\mathbf{T}_k = \begin{bmatrix} \mathbf{R}_k & \mathbf{t}_k \\ \mathbf{0} & 1 \end{bmatrix}. \tag{4}$$

A set of camera poses (provided by the EMT system) on their own do not give us enough information to create a mosaic. To that end, a set of homographies is necessary, requiring also the planar structure modeling the scene. We now detail how to obtain a set of homographies from both poses and the surface plane. Provided that only camera poses are measured, we need to establish a link between these and imagery.

Let us consider a virtual camera with its virtual image plane located in the origin of coordinates, which we use as a reference, whose image plane space is Ω_M. We can then link the spaces of the images obtained by the cameras, and their respective motions provided that the imaged plane is known. Since it is so, there is a homography $\mathbf{H}$ defined that propagates any point in the virtual image to any other image through the following equation, that for convenience we define as $g(\cdot)$

$$\mathbf{H} = g(\mathbf{x}, \mathbf{v}) = \mathbf{K}(\mathbf{R} - \mathbf{t}\mathbf{v}^T)\mathbf{K}^{-1}, \tag{5}$$

where $\mathbf{K}$ is the intrinsic camera calibration matrix, and $\mathbf{v}$ corresponds to the unit normal vector to the imaged surface $\mathbf{n}$ observed from the origin of coordinates, divided by the distance d between origin of coordinates and the plane

$$\mathbf{v} = \frac{\mathbf{n}}{d}. \tag{6}$$

We use this relation in order to relate the visual content in the images with EMT readings of the camera poses. The complete derivation of Eq. (5) can be found in Ref. 16.

In the following work, Ref. 6 introduced a probabilistic graphical model that infers the set of poses $\mathcal{X}$ and planar structure $\mathbf{v}$ by solving the minimization problem

$$(\hat{\mathcal{X}}, \hat{\mathrm{v}}) = \underset{(\mathcal{X}, \mathrm{v})}{\arg\min}(C_v + C_{\mathrm{EMT}}), \tag{7}$$

where $\hat{\mathcal{X}}$, $\hat{\mathbf{v}}$ are the set of estimated camera poses and plane, respectively, the cost C_v represents the sum of all reprojection errors between matching landmarks with pairs of indices $\{l, m\} \in \mathcal{L}$ found between all images where the homographies $\mathbf{H}_l$, $\mathbf{H}_m$ are obtained using the three components; the two camera poses $\mathbf{x}_l$, $\mathbf{x}_m$, and plane $\mathbf{v}$ as in Eq. (5). The EMT information is incorporated in the second cost C_{EMT} as

$$C_v = \sum_{\{l,m\} \in \mathcal{L}} \sum_{i=1}^{N_{l,m}} \frac{1}{\sigma_v^2} \| \mathbf{p}_l^i - f(\mathbf{H}_l \mathbf{H}_m^{-1} \tilde{\mathbf{p}}_l^i) \|_2^2, \tag{8}$$

$$C_{\mathrm{EMT}} = \sum_{k=1}^{N} (\mathbf{z}_k - \mathbf{x}_k)^T \Sigma_{\mathrm{EMT}}^{-1} (\mathbf{z}_k - \mathbf{x}_k). \tag{9}$$

This prevents the poses to be estimated from drifting provided that the EMT measurements have a common reference, which do not allow drift to separate the cameras from the measured position. In other words, the solution obtained as a result of Eq. (7) reflects a guided estimation of the parameters using the EMT system.

2.2 *Pruning Strategies for Sequential Bundle Adjustment*

In this section, we present our two pruning strategies: (i) the use of the EMT system to identify and discard nonoverlapping frames for which no matching should be attempted, and (ii) the introduction of the concept of superframe; an extension of a typical frame that allows for more efficient mosaicking schemes. We present two pipelines that increasingly incorporate the contributions mentioned above.

2.2.1 *Frame pruning using the EMT system*

Given the small field of view in fetoscopy and the relatively large area to explore, only a small subset of images shares visual content. Global alignment schemes typically attempt to find correspondences between each pair of images, although in many cases such as ours, the majority of image pairs do not overlap spatially and thus cannot be matched. A given image will typically only match to a small spatially adjacent subset. For this reason, we propose to use the information of an EMT system to find plausible image candidates, bypassing the computation of unnecessary failed matching attempts and reducing, thereby the computational complexity of bundle adjustment.

At a given time instant k, we have the current EMT measurement z_k and the previous estimation of the plane v_{k-1} by solving the nonlinear optimization problem posed by Eq. (7). Therefore, we can compute a noisy homography using Eq. (5) as

$$\mathbf{H}_k^{\mathrm{EMT}} = g(\mathbf{z}_k, \mathbf{v}_{k-1}). \tag{10}$$

This estimation of the homography, even if not very accurate, can be used to decide whether two images are likely to overlap, thus allowing to filter out spatially unreasonable candidates. To do so, once the homography is obtained, we project the corners of the bounding box containing the circular fetoscopic region of interest through $\mathbf{H}_k^{\mathrm{EMT}}$ onto the mosaic space as

$$\tilde{\mathbf{p}}_{k,c}^{\Omega_M} = \mathbf{H}_k^{\mathrm{EMT}} \tilde{\mathbf{p}}_{k,c}, \tag{11}$$

where the points $\tilde{\mathbf{p}}_{k,c}$ and $\tilde{\mathbf{p}}_{k,c}^{\Omega_M}$ represented in homogeneous coordinates, are the original and propagated image corners to the mosaic space Ω_M, respectively, where c denotes the corner index. Then, the overlap between the current and previous frames can be easily and efficiently obtained. Kekec et al.[8] showed how this problem can be seen as a simple convex polygon intersection problem. The estimation of the overlap given two convex polygons can be efficiently solved using the separating axis theorem (SAT);[8] a theorem from computer graphics that states that if a straight line can be drawn between two convex polygons, then the polygons do not overlap.

2.2.2 *Pairwise compression*

A practical problem in mosaicking[12] algorithms is that hundreds of correspondences are typically stored and used in the optimization procedure per image as can be seen in Eq. (9). Instead, we parametrize the pairwise relation as five equivalent correspondences in the following way: (i) after the pairwise homography estimation using RANSAC simultaneously with the correspondences acquisition,[17] we take the bounding box that includes all the interest points in the image. In particular we take the top-centered, bottom-centered, left, and right point locations as well as the center of the bounding box to account for the spread of these points in the original image space. (ii) We propagate them using the estimated homography that relates both images to obtain the second set of points, which completes the collection of correspondences. These correspondences will then be kept for the optimization, reducing the run-time of the cost function greatly.

The more correspondences we acquire, the more correspondences must be taken into account in the cost function. This might not be a problem at the beginning of the sequence, but as the number of correspondences increases, the computational cost increases as well, hindering online operation. We now introduce an additional strategy that generalizes a regular frame in a more efficient representation for mosaicking.

Figure 2(a) shows the proposed pipeline proposed in this section while Fig. 2(b) shows the diagram of the pipeline proposed in Sec. 2.2.3. In both pipelines, the pairwise compression mentioned in Sec. 2.2.2 is applied. Contributions to the mosaicking pipeline are outlined with a blue border.

2.2.3 *Superframe representation*

The superframe representation is a generalization of a frame that incorporates one or more frames. The main idea is to partition the image set $\mathcal{I} = \{\mathbf{I}_k\}_{k=1}^{K}$ into subsets of W images grouped into N superframes with $K = NW$. Since it is a generalization of an image, the superframe can be incorporated into the standard bundle adjustment pipeline reducing its computational burden drastically. Let us formally define the concept of superframe.

We define a superframe as a representation of a subset of W frames $\mathcal{I}_i = \{\mathbf{I}_k\}_{k\in\mathcal{K}_i}$ that encodes the most salient information of the region observed by all images in the superframe which are indexed by $\mathcal{K}_i = \{k \in \mathbb{Z} | LF_i - u \leq k \leq LF_i + u\}$. A lead image $\mathbf{I}_{LF_i}$ with index $LF_i = (i-1)W + u + 1$ is defined as the central image in the superframe of window of size $W = 2u + 1$, u being an integer. If the lead image was taken in isolation, then this would be equivalent to the concept of keyframes.[12] In contrast, the superframe uses information of all the images in the window.

Analogously to the standard pipeline, interest point locations $\mathbf{P}_k$ and descriptors $\mathbf{D}_k$ are extracted for all frames within each superframe. We propose to use the lead image space as a common space where all interest points in the superframe lay, defining the superframe as $\mathcal{S}_i = \{\mathcal{P}_{S_i}, \mathcal{D}_{S_i}\}$. To propagate the locations of interest points into this common space, we define $\tilde{\mathbf{P}}_k \in \mathbb{R}^{3\times N_k}$ as a matrix containing all point locations in homogeneous coordinates, where N_k is the number of interest points found in image k. If expressed this way, the points can be propagated onto the lead image space as

$$\mathcal{P}_{S_i} = \{f(\mathbf{H}_{LF_i,k}\tilde{\mathbf{P}}_k)\}_{k\in\mathcal{K}_i}, \tag{12}$$

while the descriptors are grouped as

$$\mathcal{D}_{S_i} = \{\mathbf{D}_k\}_{k\in\mathcal{K}_i}, \tag{13}$$

where the homography $\mathbf{H}_{LF_i,k}$ is the homography that relates the frame k within the superframe with the lead frame LF_i. This homography is obtained by first running a local, small-scale visual bundle adjustment with all W images following the methodology described in Sec. 2.1. Within the superframe, the aligned homographies are then considered fixed in the rest of the pipeline. Figure 3(a) depicts the process of the creation of a superframe, and Fig. 3(b) shows the differences in matching between two images (top) and two superframes (bottom).

2.2.4 *Superframe in the mosaicking pipeline*

To integrate the superframe into the mosaicking pipeline, we extend the probabilistic framework[6] summarized in Sec. 2.1. However, now superframes replace single frames. To highlight why this is possible, let us describe the simple scenario where two superframes are created. In that case, their descriptors can

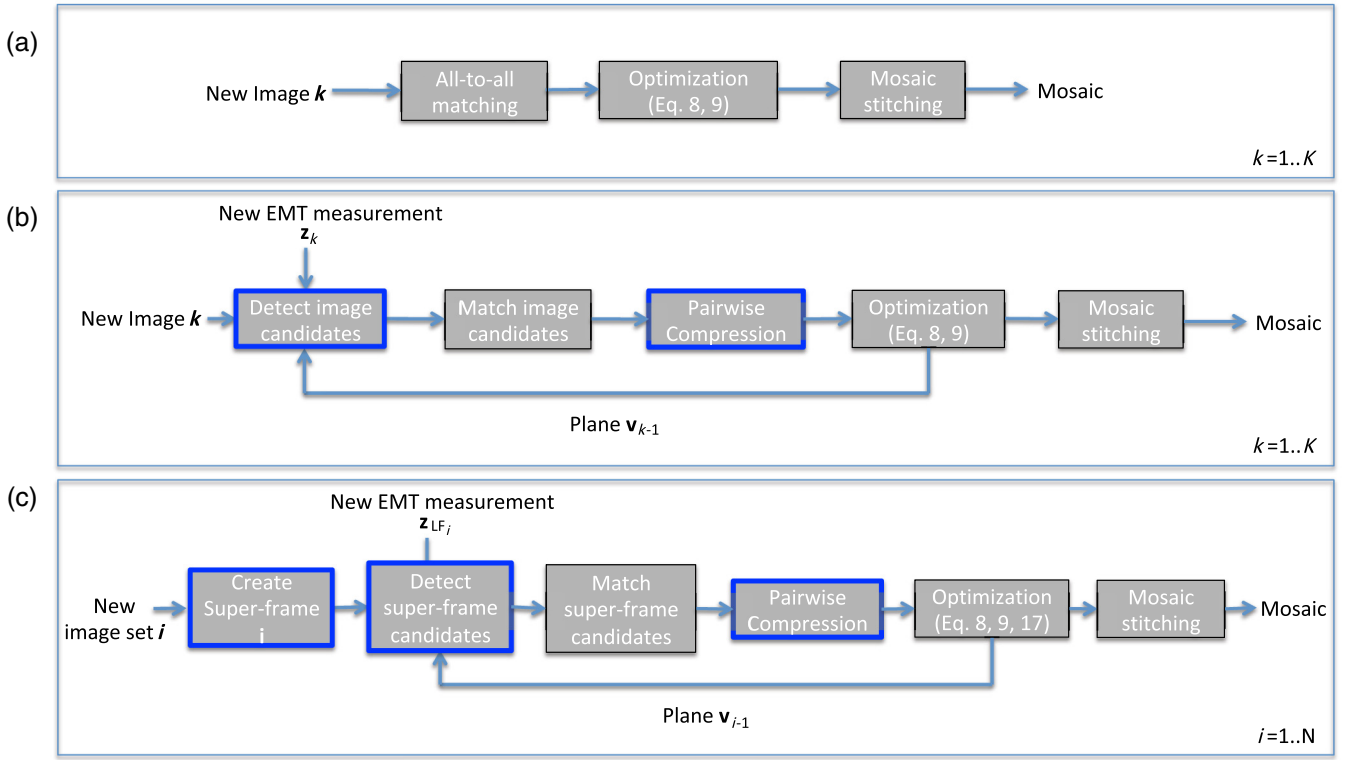

Fig. 2 (a) Pipeline proposed in Sec. 2.2.1 using the EMT system to filter out incoherent matching attempts. (b) Pipeline proposed in Sec. 2.2.3 using the superframe. In both pipelines, the pairwise compression mentioned in Sec. 2.2.2 is applied. Contributions to the mosaicking pipeline are outlined with a blue border.

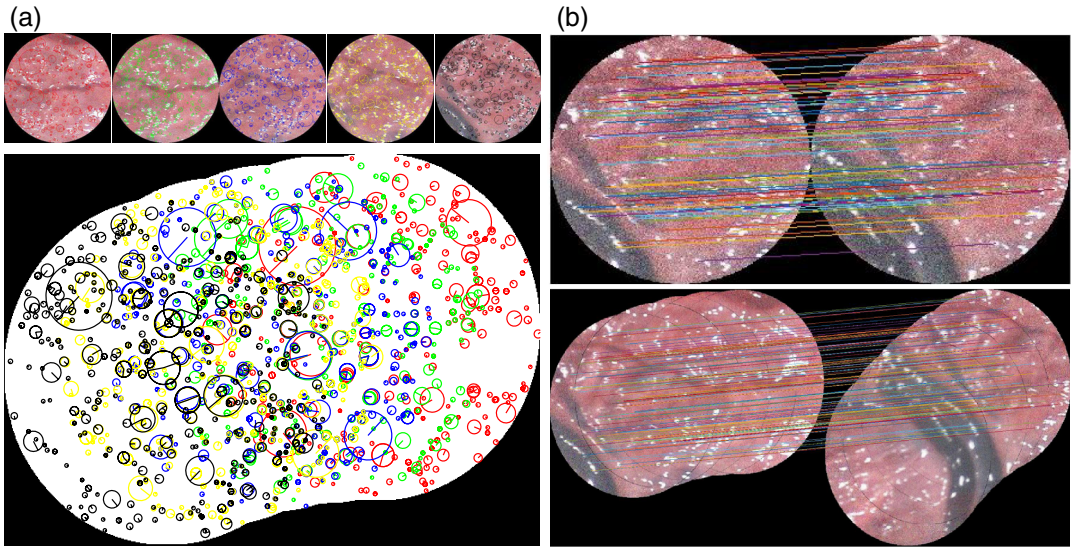

Fig. 3 (a) Creation of a superframe. From top to bottom: W images are registered using bundle adjustment[10] and their interest points are propagated to the space of the lead frame i, forming a superframe. (b) Matching of two frames (85 inliers) and two superframes (165 inliers).

be matched directly since the input to the matching algorithm is two sets of descriptors. From there, the matching process is analogous to the one described in the PM in Sec. 2.1. Figure 2(c) shows the diagram of the pipeline using the superframe, marking in blue the contributions in the pipeline. In applying directly,[6] only one EMT measurement is assumed to be associated with each frame. Therefore, there would be $W-1$ EMT measurements unused per superframe. Instead, we propose to modify the pipeline to include all EMT measurements in the window W for better placement of the superframe in the mosaic.

These measurements can be used to obtain homographies $\mathbf{H}_k^{\mathrm{EMT}}$ using Eq. (5) and the previously estimated $\mathbf{v}_{i-1}$ resulting from Eq. (7). Figure 4 shows the nomenclature and indexing,

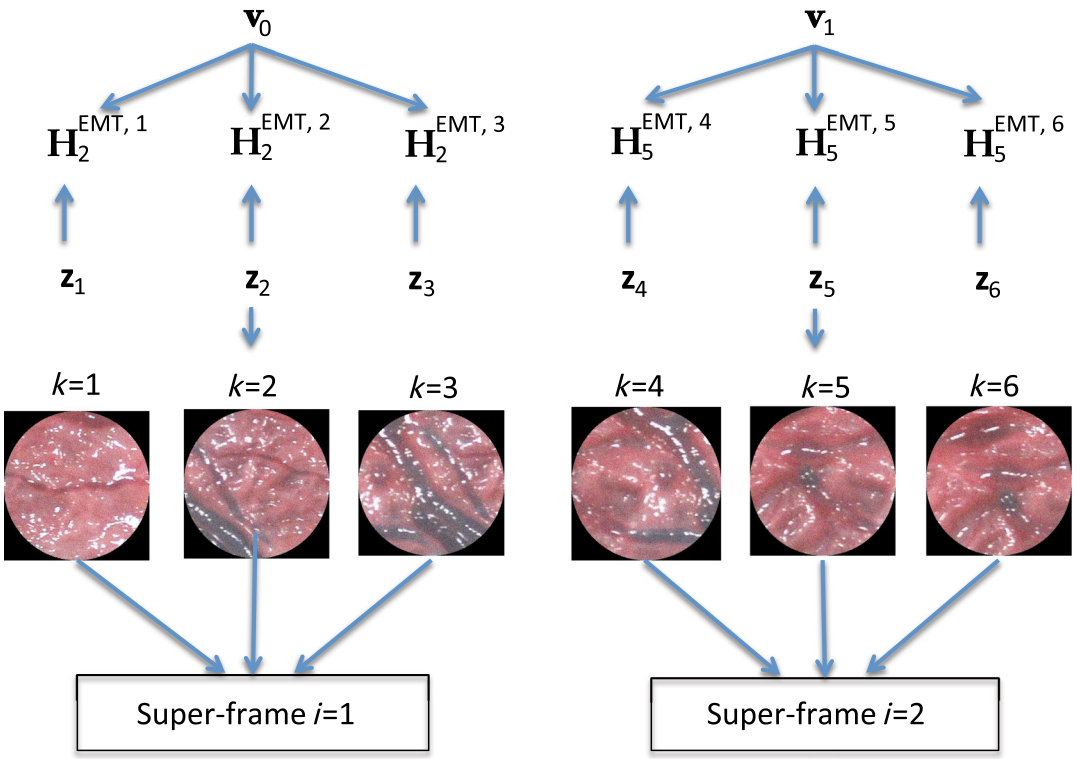

Fig. 4 A superframe indexed by i is composed of a set of single frames indexed by k. Each homography $\mathbf{H}_{LF_i}^{\text{EMT},k}$ is created from the EMT measurement $\mathbf{z}_k$ and the previous plane estimation $\mathbf{v}_{i-1}$.

displaying each homography composed using the EMT system. Using the EMT measurements and the fixed homographies $\mathbf{H}_{LF_i,k}$ found in the initial bundle adjustment to build the superframe,, we can obtain measurements of the lead homography $\mathbf{H}_{LF_i}^{\text{EMT},k}$ coming from each of the EMT measurements in a given superframe as

$$\mathbf{H}_{LF_i}^{\text{EMT},k} = \mathbf{H}_{LF_i,k}\mathbf{H}_k^{\text{EMT}} \quad k \in \mathcal{K}_i. \tag{14}$$

We consider these homographies as measurements of the true homography that places the lead image into the mosaic space, and formulate the problem as

$$(\hat{\mathcal{X}}, \hat{\mathbf{v}}) = \arg\min_{(\mathcal{X},\mathbf{v})}(C_{\text{EMT}} + \lambda C_v + \lambda' C_H), \tag{15}$$

where λ and λ' are weights. The two first costs are precisely the costs defined before in Eqs. (9) and (8), and the proposed additional term for the global alignment cost function C_H is defined as

$$\epsilon_k(c,i) = f(\mathbf{H}_{LFi}^{\text{EMT},k-1}\tilde{\mathbf{p}}_c) - f[g(\mathbf{x}_k, \mathbf{v}_{i-1})^{-1}\tilde{\mathbf{p}}_c], \tag{16}$$

$$C_H = \frac{1}{NWQ}\sum_{i=1}^{N}\sum_{k\in\mathcal{K}_i}\sum_{c=1}^{Q}\epsilon_k^T(c,i)\sum_{hVis}^{-1}\epsilon_k(c,i), \tag{17}$$

where $\mathbf{p}_c$ represents Q control points. Specifically, we used the four corners and the center point of an image as control points. Setting $\sum_{hVis}$ is not trivial since it depends on the accuracy of the EMT system and previous estimation of the normal vector. Therefore, we use it conservatively setting a standard deviation of 10 pixels.

Once all homographies from the mosaic space to the lead frames of each superframe are obtained, one must place the superframes into the mosaic space. To do that, each one of the images within the superframe has to be placed in the mosaic space. Therefore, we must obtain individual homographies $\mathbf{H}_k$ for each frame in the sequence that relates it with the mosaic space. One must note, however, that when running bundle adjustment on the superframes, the reference is in the space of the first lead frame, which corresponds to index LF_1. For simplicity, we denote the estimated homographies as $\mathbf{H}_{LF_i}$, skipping the second subindex, which would be LF_1. However, one must keep in mind that they are actually from the first lead frame LF_1 to LF_i as $\mathbf{H}_{LF_i,LF_1}$. Therefore, to compose a mosaic in the space of the first image, not the first superframe, one must use

$$\mathbf{H}_k = \mathbf{H}_{k,LF_i}\mathbf{H}_{LF_i,LF_1}\mathbf{H}_{LF_1,1}. \tag{18}$$

Once the absolute set of homographies is obtained, one can project the images onto the mosaic space to be blended.

2.2.5 *Efficient implementation of the pipeline using superframes*

The main idea of the superframe is to summarize many frames into a region by first performing a local bundle adjustment of W frames, each with a complexity of $\mathcal{O}(W^2)$. This entails all-to-all matching of all images within the superframe. On the other side, the complexity of a full bundle adjustment (FBA) is quadratic in the number of frames $\mathcal{O}(K^2)$. By using superframes instead of single frames, one allows reducing the number of effective images used in a general bundle adjustment. This procedure now takes into account $\frac{K}{W}$ superframes, so the total complexity is only $\mathcal{O}(W^2N + \frac{K^2}{W})$ where $K = NW$. The first term accounts for the

small bundle adjustments of size W to build each superframe and then the second term accounts for the complete bundle adjustment of the superframe. For example, in a sequence of $K = 100$ frames, where the window $W = 5$, we only have $N = 20$ superframes. Therefore, the construction of the superframes would take $\sim 500\tau + 400\tau'$ whereas a single FBA would take $10{,}000\tau$, where τ and τ' are timings related to interest point detection and matching, which differ for single frame and superframe, depending on the number of frames W used to create each superframe.

The number of interest points contained in a superframe would be the sum of the number of interest points detected in each image. As W increases, this number can quickly become too large and prohibitive for matching. To reduce the number of interest points in the superframe drastically, we put two strategies in place. First, we greedily select only interest points resulting in being inlier matches within every frame in the creation of the superframe. Intuitively, interest points that have already been matched should have a higher probability of being matched in the superframe compared to other points that might not be as well described. The second strategy is to keep a precomputed a KD-tree,[23] implementing an approximate nearest neighbors strategy for efficient matching in the creation of the superframe.

In addition, the matching candidate selection strategy mentioned in Sec. 2.2.1 is adapted to be used with superframes instead of single frames. To apply it, only the meaning of overlap for two superframes has to be redefined: two superframes overlap if any of their single frames overlap. In this way, we can incorporate the computational advantage of using superframes that leverage the EMT system to filter out spatially incoherent matching candidates.

3 Experiments

In this section, we present our experimental suite to demonstrate the computational advantage of the proposed pruning strategies compared against a baseline implementation of bundle adjustment. We start by presenting the datasets, algorithms, metrics, and implementation details.

3.1 *Datasets*

We generated two datasets formed by a set of images and EMT data; synthetic (SYN, 273 frames, 373×378 px) and a phantom based (PHB, 701 frames, 780×781 px) datasets. SYN is a translation-based raster scan generated by simulating ground truth camera motions, from which EMT measurements are generated by applying Gaussian noise,[6] and for which the images were extracted as projections of the scene observed by the cameras. In that case, the scene consisted of a planar, high-resolution digital image. PHB was recorded by imaging a printed image of a placenta. The setup consisted of a camera head IMAGE1 H3-Z SPIES mounted on a 3-mm straight scope 26007 AA 0° (Karl Storz Endoskope, Tuttlingen, Germany), an EMT system NDI Aurora with a planar field generator, and a Mini 6 DoF sensor. A collection of homographies by semiautomated registration of each fetoscopic image to the original image of the placenta, where a landmark-based approach was used to align the images after the initial manual alignment.

3.2 *Algorithms*

As a baseline, we compare our algorithms to both the established FBA as well as the standard PM approach, which is very fast but accumulates drift. We apply the matching compression strategy presented in Sec. 2.2.2 as well as the use of the EMT system to discard inconsistent matching attempts (SAT). Our second contribution is the introduction of the superframe (SF) and its incorporation to the mosaicking pipeline. By default, we have chosen $W = 5$. We explicitly name SF(W) to a run where W corresponds to the size of the window used in the superframe.

3.3 *Metrics and Implementation Details*

We compare the homography $\mathbf{H}$ and the ground truth homography $\mathbf{G}$ by computing the mean residual error ϵ_k of a projected grid of N_g points ρ_i from the image space to the mosaic space. Then, the error ϵ between two mosaics is computed as the mean of individual errors ϵ_k. This error is defined in the mosaic space, which is the space in which the final composition is performed. Note that we use ϵ^k with superindex k to refer to the residual error between two sequential mosaics at each time instant whereas ϵ_k with subindex k refers to the error of the alignment of a single image k. To make this metric independent to the choice of reference space, we compute the average of the errors using all images as a reference, as in

$$\epsilon_k = \frac{1}{KN_g} \sum_{j=1}^{K} \sum_{i=1}^{N_g} \left\| f(\mathbf{H}_{j,k}^{-1} \rho_i) - f(\mathbf{G}_{j,k}^{-1} \rho_i) \right\|_2. \tag{19}$$

In terms of feature-based method, we used SURF.[20] The matching was performed using fast approximate nearest neighbors.[23] The fetoscope was precalibrated using the Matlab Camera Calibration Toolbox. We also precomputed and applied the Hand-eye Calibration[24] matrix from a sequence of images of a checkerboard as well as synchronized sensor poses. The window size of the superframe was determined from the trade-off analysis as $W = 5$, which is presented later in Sec. 3.3. In addition, given that superframes contain many more correspondences than actual frames, a kd-tree[23] can be trained and stored per superframe. Then, fast approximate nearest neighbor can be used to accelerate the matching between superframes. The experiments have been performed in an Ubuntu 16.04 with Intel Core I7 at 2.5 GHz and 6 GB of memory.

Next, we evaluate sequential bundle adjustment and evidence the decrease in computational time that the proposed strategies provide.

3.4 *Speedup: Correspondence Acquisition*

In this section, we analyze the computational cost of a sequential bundle adjustment in an online setting. More precisely, we focus on correspondence acquisition.

In Fig. 5, we display the run-time of correspondences acquisition in the three algorithms for both datasets. As expected, SAT is faster than FBA in both cases, and SF shows the best in terms of computational cost. In the case of the phantom-based dataset, we can observe that SAT is not necessarily faster all the time. This is due to the trajectory that the camera has followed in this particular dataset. If the overlap between images is too large, then the algorithm suggests all images as possible candidates.

To further evidence where the computational savings have happened spatially, we analyze the percentage of potential image candidates that have been selected using an occupation matrix. This matrix shows whether there has been a match from

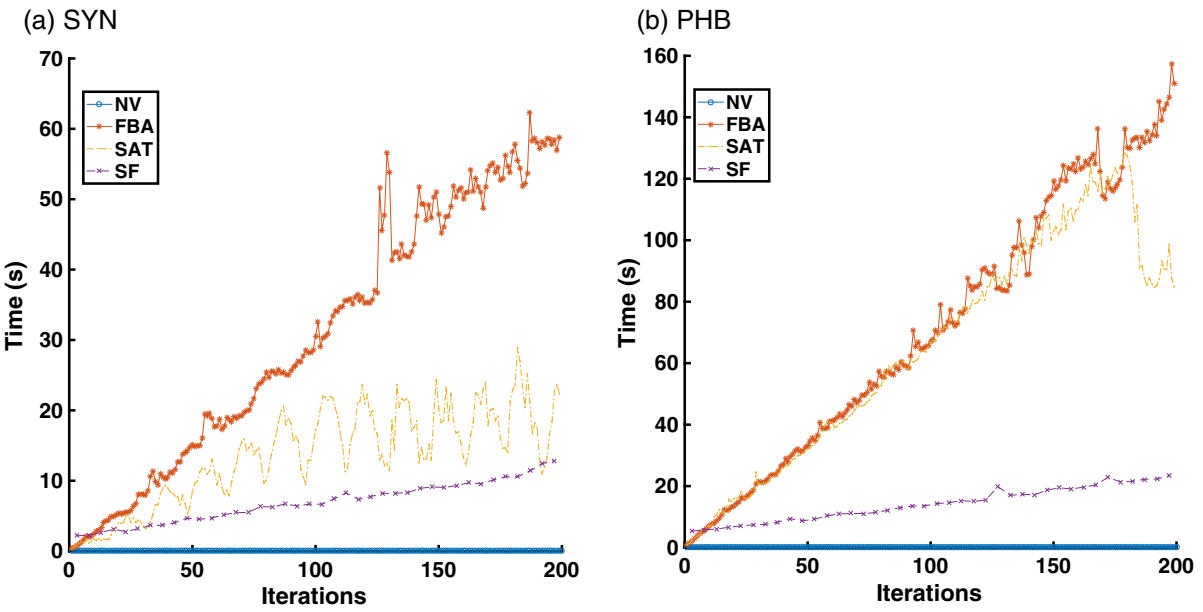

Fig. 5 Runtime of the correspondences acquisition with respect to the number of iterations for both SYN and PHB datasets.

image i (corresponding to row i) to image j (corresponding to column j). For a more informative representation, we have color-coded the matrices. Out of all image matches that would have been obtained using an FBA (white), only a portion of them would be selected by SAT (green). Due to the use of symmetrical matching, we only display the upper right triangle in the occupancy matrix. Figure 6 shows that, in the case of SYN (a), 100% of the proposed candidates are correct, and no matches have been missed, saving a total of 54.2% matching attempts. As mentioned before, the path followed by the camera in the PHB [Fig. 6(b)] makes saving matching attempts more complicated. 100% of the proposed candidates are correct; however, there is only a total saving of 5.2% attempts, which happened to be toward the end of the sequence.

We now study the run-time of the nonlinear optimization as the second bottleneck for bundle adjustment to reach sequential operation.

3.5 *Speedup: Nonlinear Optimization*

The computational cost of the nonlinear optimization is due to the run-time of the cost functions in Eqs. (9) and (8) and the number of iterations to convergence. In contrast to standard bundle adjustment where we do not have any initial estimate *a priori*, subsequent estimations can be used as an initial estimate in a sequential version of bundle adjustment, providing a speedup at each iteration due to the proximity to the minimum. Although this is a desirable property, the question that might

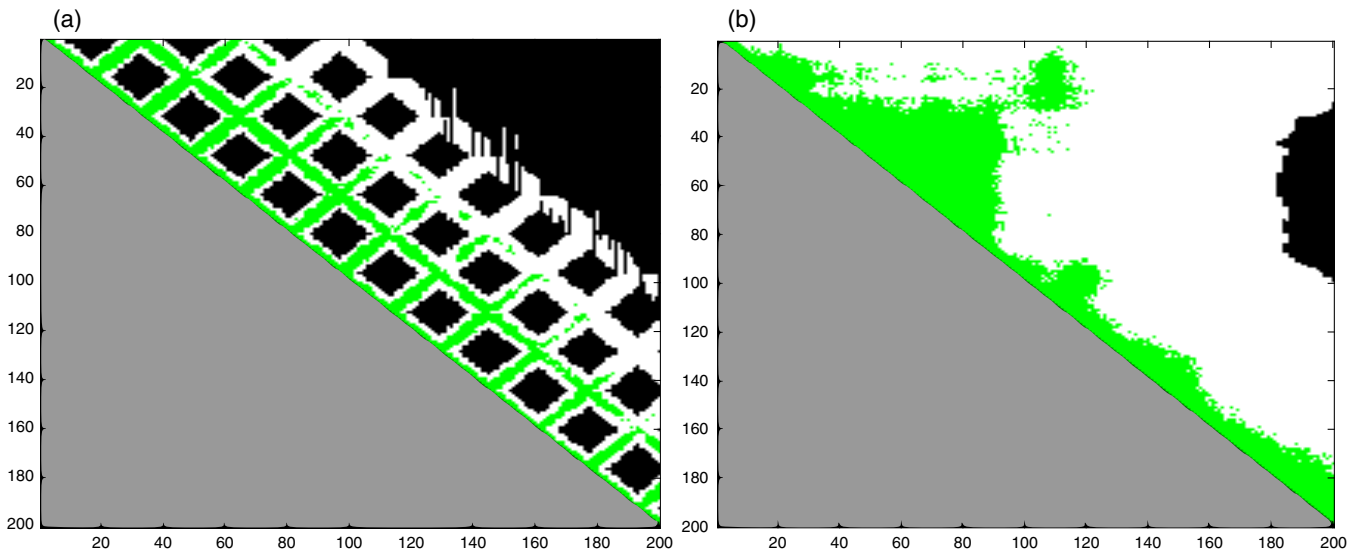

Fig. 6 Occupancy matrix. The pixel (i,j) shows a color-coded outcome of the matching process. Potential matching candidates are in white. Proposed candidates that have been matched are in green. Black pixels reflect that no candidate was selected but there was no real match. Occupation matrices in (a) the synthetic dataset (SYN) and (b) the phantom-based dataset (PHB).

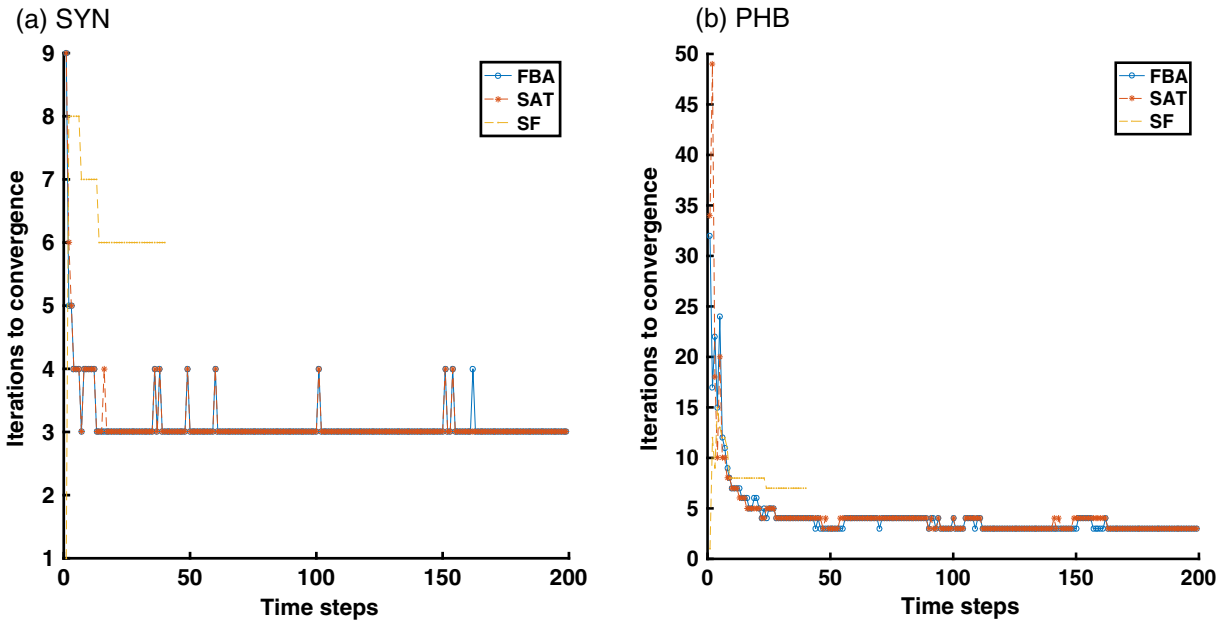

Fig. 7 Number of iterations of the nonlinear optimization procedure for FBA, SAT, and SF.

prevent desirable update rates is whether the number of iterations to convergence grows with time.

In Fig. 7, we show the number of iterations to convergence in both SYN and PHB datasets.

The fact that the approaches show a constant tendency over time is crucial since it indicates that the number of iterations in the nonlinear optimization does not seem to be a bottleneck for offline operation. Given the textureless nature of fetoscopic images, using only a subset of all frames is not a very robust option. Instead, we would like to use the redundancy in the complete video sequence. Figure 8 shows the optimization times for FBA, SAT, and SF. It can be seen how if compression is used (SAT), the slope decreases with the number of iterations. Now that we have demonstrated the decrease in computational burden in the proposed algorithms, we need to evaluate their accuracy and show that despite the improvement in speed, there is no significant loss in accuracy.

3.6 *Efficiency*

In this section, we analyze the performance of the proposed algorithms for classical pipelines. In particular, we computed the error in each sequential mosaic for NV and FBA as a baseline, comparing it to SAT, and SF with different window sizes $W = 1,3,5,7,9$ in SYN and PHB datasets, respectively. Figure 9 shows the errors for each of the methods. The error in NV spikes from the beginning. This is well known due that there is no global correction and therefore, drift is accumulated. While

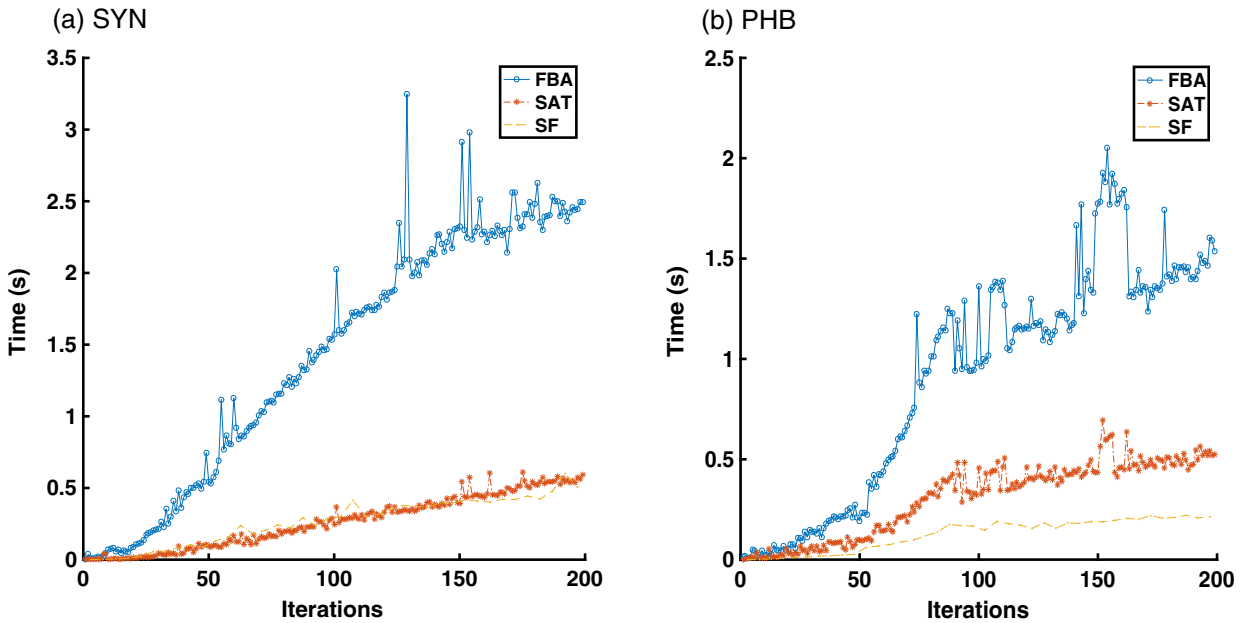

Fig. 8 Runtime of the nonlinear optimization for FBA, SAT, and SF.

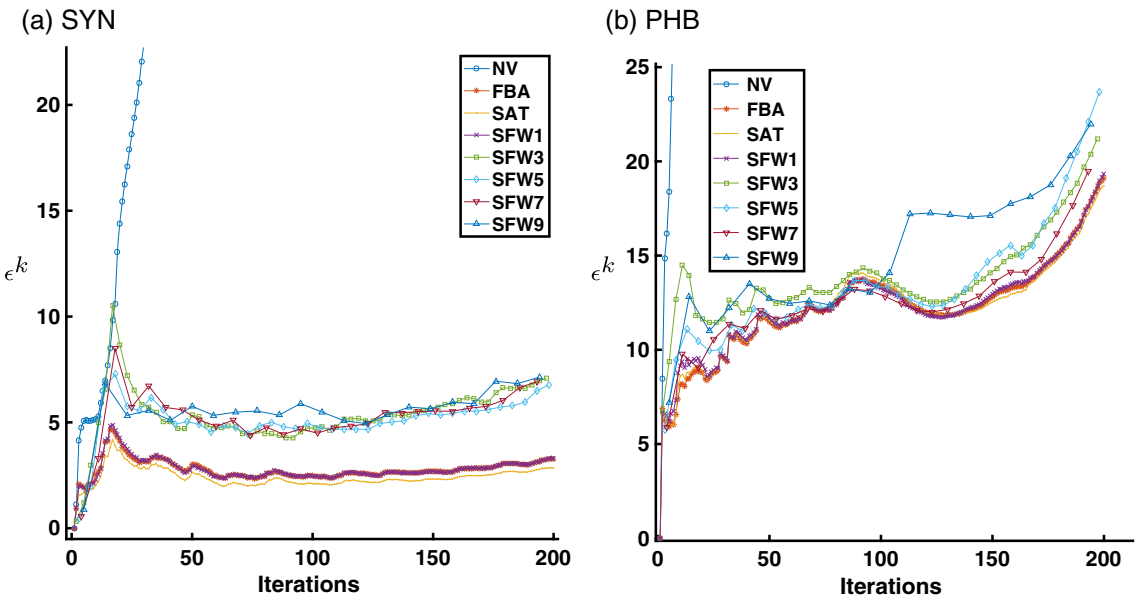

Fig. 9 Error in NV, FBA, SAT, and SF with a window size of 1, 3, 5, 7, and 9 for SYN and PHB.

SF results in the best performance regarding computational advantage, the decrease in accuracy for all the window sizes except $W = 1$ for which the behavior should be very similar to SAT. When the window increases, the accuracy of the SF remains stable.

Since the data reflects this trade-off between computational power and accuracy, we propose to analyze this trade-off with the graph in Fig. 10. This graph shows the error in the y axis and the time per iteration in the x axis. We have plotted each algorithm in a different color. In here, temporal evolution is not shown; however, one can observe the temporal trail by looking at the lines connecting the dots, taking into account that first iterations are the fastest. For an efficient algorithm, we would like the error and computational time to be low. Therefore, we identify the most efficient algorithm as the one whose results lie in the bottom left corner. The same tendency can be seen for both datasets; despite being efficient, the NV drifts in large measure, making the algorithm not worth pursuing. The FBA is not a good option concerning the trade-off since the computational time is too large, yet it does keep the accuracy very low. SAT performs similarly, albeit much more efficiently than FBA. On top of that, all versions of SF perform more efficiently than SAT. In particular, we see that as we increase the window size, the efficiency starts to be better until $W = 5$. The selection of the best algorithm is not trivial since $W = 3$ shows a faster start yet slower end than $W = 5$ in both datasets due to the need for matching more superframes. After this, efficiency starts to fall systematically. We believe that the increase of computational time that takes to perform the first bundle adjustment causes this effect, growing with W. Since many of the operations in the creation of the superframe could be performed in parallel, we expect slightly better results for a parallelized version of the approach.

In Figs. 11 and 12, we show mosaics for SYN and PHB datasets, respectively, for the NV, FBA, and SFW5. While the figure shows some steps of the sequential estimation of the mosaic, we encourage the reader to visualize the videos included in the supplementary material for a more thorough visualization of the results.

4 Discussion

Bundle adjustment is in general slow due to the mentioned reasons. In this work, we propose the use of the EMT system to identify potential matching candidates, avoiding unnecessary matching attempts. However, the computational savings will greatly depend on the motion of the camera. For example, if the camera remains in the same position, then the EMT system will determine that all the frames are candidates being equivalent to an all-to-all matching scenario. On the contrary, if the camera moves, the EMT system is able to identify a reduced set of potential candidates. A simple solution to this problem could be set up a preprocessing step of filtering out the frames that overlap more than a certain area threshold.

The accuracy of the EMT is limited. Discarding nonmatching candidates is a process that can be made more loose or aggressive by just increasing or decreasing the size of the images when determining overlapping frames. This will depend on the accuracy of the EMT when projected to the image space, i.e., the noise on the EMT system, but also the camera parameters, and the placement of the planar object. In our datasets, results seem to suggest that taking the image size is enough to discard incoherent matching candidates.

However, the estimation of the overlap is dependent on having a reliable normal estimation. When the camera has not observed enough scene, the estimation of the normal vector is less accurate due to the lack of depth perception in a small baseline. When more parts of the scene are observed, the normal vector converges, and only a few constant number of iterations are necessary for convergence in subsequent iterations. In fact, the results suggest that only a constant number of iterations are enough for the nonlinear optimization to converge. This is a significant result that reveals that the number of steps to convergence is not a problem for achieving clinically feasible update rates. The obvious exception to this point is if there are outliers in an image. If so, then the cost function does not reflect the problem to solve and as a consequence, the number of iterations to converge grows, not reaching the desired minimum. Therefore, it is essential to obtain an outlier-free set of correspondences.

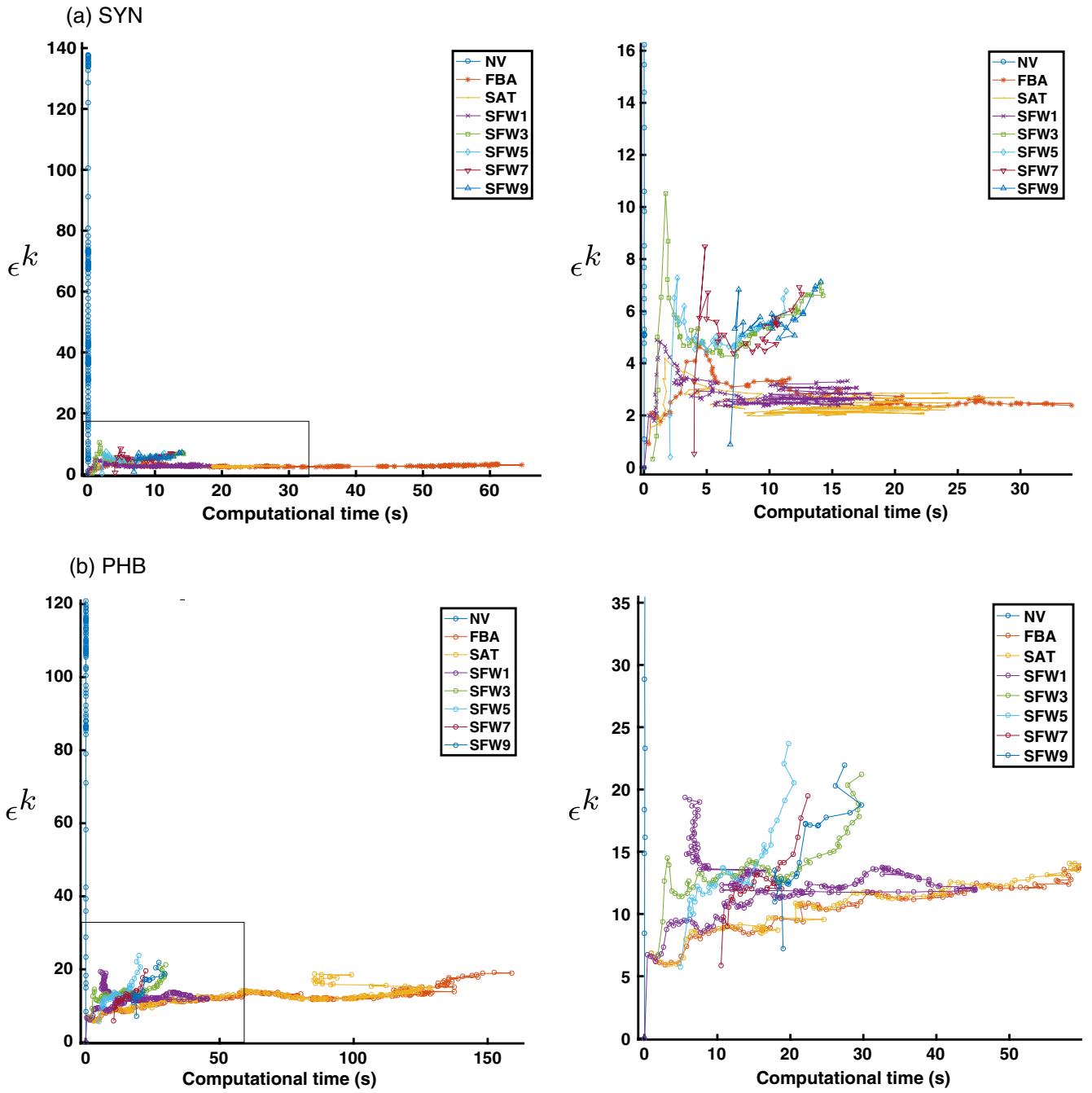

Fig. 10 The trade-off between the error ϵ^k in the y-axis and computational time in the x-axis. To the left, a zoomed version of the graph in the right for both (a) SYN and (b) PHB. Each color represents a different algorithm and despite the graph does not aim to show the temporal evolution, the connection between points provide an indication.

The creation of a superframe requires a first bundle adjustment. This process, even if it takes some time, is compensated by the fact that the general bundle adjustment takes only $\frac{K}{W}$ frames into account. The efficiency gain is expected, whenever, the number of superframes N is less than the number of frames K. However, errors in the local bundle adjustment impact in the second, global, bundle adjustment since it is only optimizing for the lead frame in the superframe. If there is an error, one of the homographies of the superframe, then the interest points will be erroneously placed within the space of the lead frame in the superframe. Then, a geometrical validation such as RANSAC could filter them out for being geometrically inconsistent with a homographic model. This results in a trade-off between the accuracy and the window size W. A second trade-off is that concerning the window size concerning efficiency. When the first bundle adjustment is large, then the number of efficient superframes to optimize in the global bundle adjustment is smaller and vice-versa. However, the larger is the superframe, the more rigid the system is, and this makes it more prone to alignment errors. Also, as the window size is larger, the distance between lead frames also increases and matching becomes more complicated.

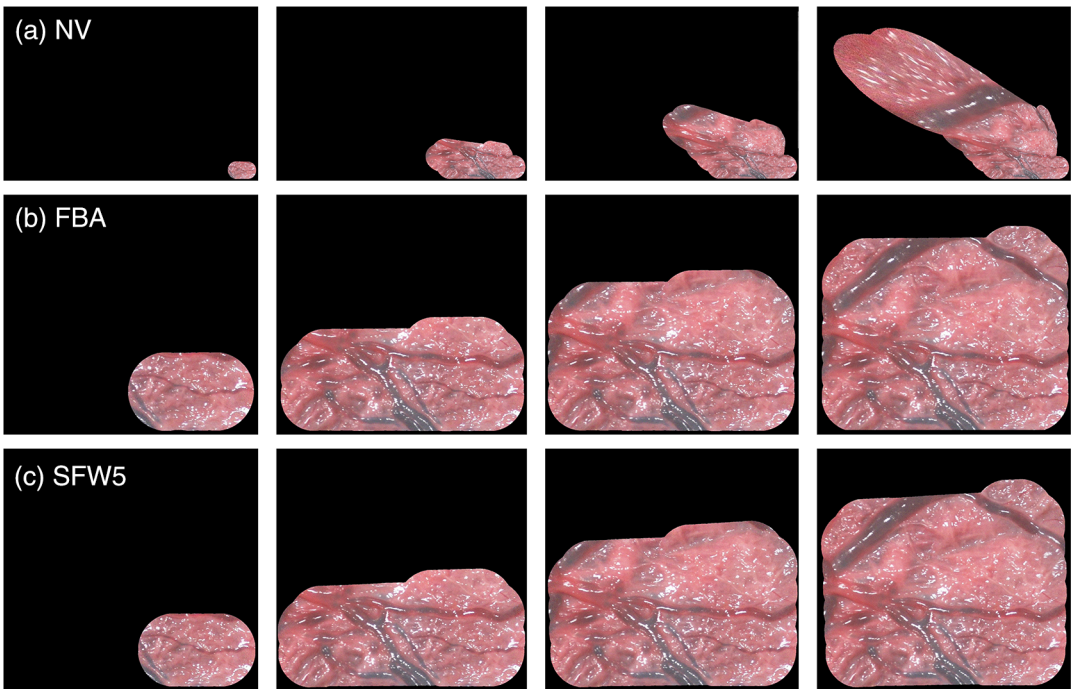

Fig. 11 Sequential screenshots of the mosaic using (a) NV, (b) FBA, and (c) SFW5 in the SYN dataset (Video 1, MP4, 1426 KB [URL: https://doi.org/10.1117/1.JMI.6.3.035001.1; Video 2, MP4, 8195 KB [URL: https://doi.org/10.1117/1.JMI.6.3.035001.2; Video 3, MP4, 2447 KB [URL: https://doi.org/10.1117/1.JMI.6.3.035001.3).

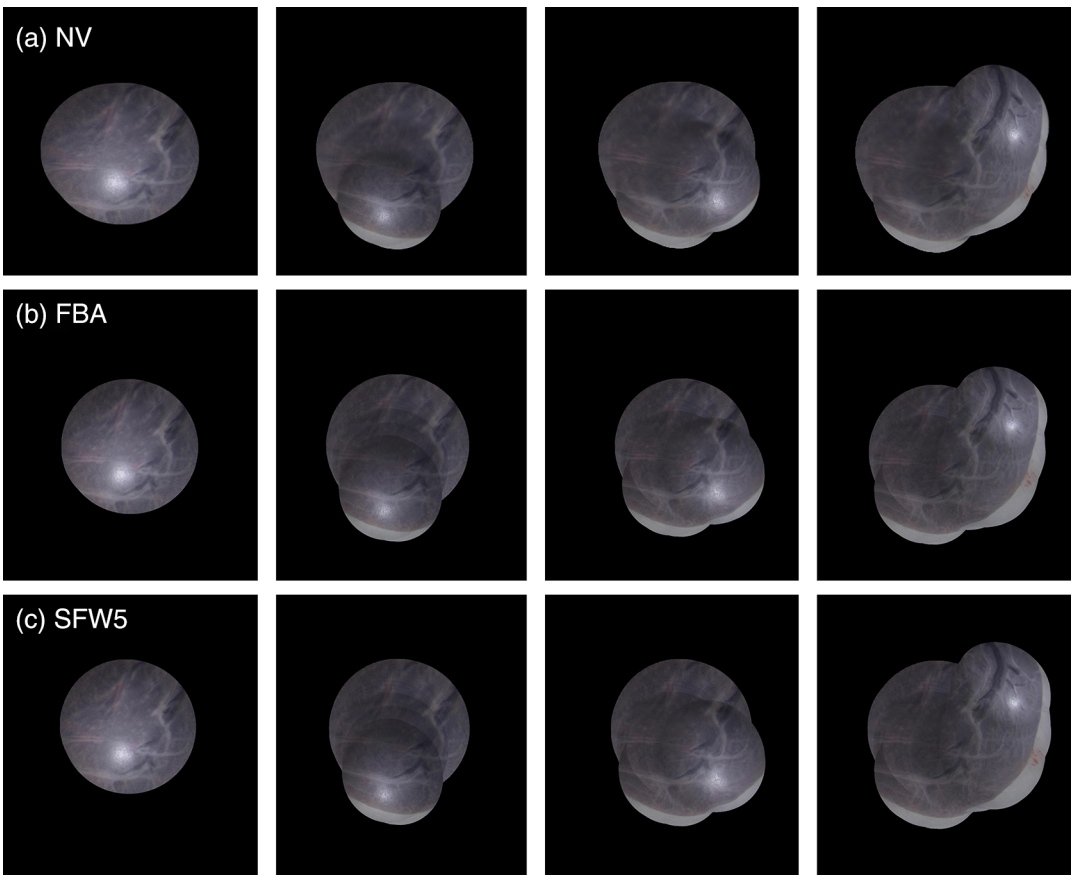

Fig. 12 Sequential screenshots of the mosaic using (a) NV, (b) FBA, and (c) SFW5 in the PHB dataset (Video 4, MP4, 2844 KB [URL: https://doi.org/10.1117/1.JMI.6.3.035001.4; Video 5, MP4, 3443 KB [URL: https://doi.org/10.1117/1.JMI.6.3.035001.5; Video 6, MP4, 811 KB [URL: https://doi.org/10.1117/1.JMI.6.3.035001.6).

In the described case where a fixed window is used, the optimal W will greatly depend on the motion on the camera making the choice of the value not trivial. We have chosen $W = 5$ as the best trade-off between accuracy and computational cost in the evaluation. An interesting fact is that the accuracy of the mosaic does not seem to decrease much between different instances of superframes even though its computational time does.

We propose a system in which single frames are not shared by the superframes, i.e., each frame index is only in one superframe. However, an interesting research line could be to inspect the impact of incorporating each index in more than one superframe. This implies that the number of frames in the nonlinear optimization procedure grows, trading an increase in the runtime of the algorithm for more robustness.

5 Conclusions

In this work, we propose two different pruning strategies allowing the sequential application of bundle adjustment for mosaicking. To make this possible, we tackle the computational bottlenecks of bundle adjustment using two different pruning strategies. First, we use the EMT system to discard a high percentage of spatially inconsistent matching attempts. Second, we introduce the concept of superframe and introduce it into the mosaicking pipeline. This concept is a generalization of an image, which can be used to reduce the computational complexity drastically in both the correspondence acquisition and optimization phases. We show a large decrease in computational complexity with respect to a standard bundle adjustment, on both synthetic and phantom-based datasets. This makes possible the use of a sequential bundle adjustment, which achieves a better compromise between efficiency and accuracy than standard approaches. The results of this work open new avenues to online operation, leading the state of the art in online mosaicking one step closer to clinical translation.

Disclosures

The authors declare no conflicts of interest.

Acknowledgments

This work was supported by the Welcome Trust (WT101957; 203148/Z/16/Z; 203145Z/16/Z) and EPSRC (NS/A000027/1; NS/A000049/1; NS/A000050/1; EP/L016478/1). Jan Deprest was being funded by the Great Ormond Street Hospital Charity.

References

1. A. Baschat et al., "Twin-to-twin transfusion syndrome (TTTS)," *J. Perinatal Med.* **39**(2), 107–112 (2011).
2. B. Münzer, K. Schoeffmann, and L. Böszörmenyi, "Content-based processing and analysis of endoscopic images and videos: a survey," *Multimedia Tools Appl.* **77**(1), 1323–1362 (2018).
3. L. Maier-Hein et al., "Optical techniques for 3d surface reconstruction in computer-assisted laparoscopic surgery," *Med. Image Anal.* **17**(8), 974–996 (2013).
4. M. Brown and D. Lowe, "Recognising panoramas," in *Int. Conf. Comput. Vision*, Vol. 3, p. 1218 (2003).
5. P. Daga et al., "Real-time mosaicing of fetoscopic videos using SIFT," *Proc. SPIE* **9786**, 97861R (2016).
6. M. Tella-Amo et al., "Probabilistic visual and electromagnetic data fusion for robust drift-free sequential mosaicking: application to fetoscopy," *J. Med. Imaging* **5**(2), 021217 (2018).
7. J. Civera et al., "Drift-free real-time sequential mosaicing," *Int. J. Comput. Vision* **81**(2), 128–137 (2009).
8. T. Kekec, A. Yildirim, and M. Unel, "A new approach to real-time mosaicing of aerial images," *Rob. Auton. Syst.* **62**(12), 1755–1767 (2014).
9. T. Vercauteren et al., "Real time autonomous video image registration for endomicroscopy: fighting the compromises," *Proc. SPIE* **6861**, 68610C (2008).
10. B. Triggs et al., "Bundle adjustment. A modern synthesis," in *Int. Workshop Vision Algorithms*, Springer, pp. 298–372 (1999).
11. P. Schroeder et al., "Closed-form solutions to multiple-view homography estimation," in *IEEE Workshop Appl. Comput. Vision (WACV)*, IEEE, pp. 650–657 (2011).
12. D. Steedly, C. Pal, and R. Szeliski, "Efficiently registering video into panoramic mosaics," in *Int. Conf. Comput. Vision*, IEEE, Vol. 2, pp. 1300–1307 (2005).
13. L. Peter et al., "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 (2018).
14. E. Garcia-Fidalgo et al., "Fast image mosaicing using incremental bags of binary words," in *IEEE Int. Conf. Rob. Autom. (ICRA)*, IEEE, pp. 1174–1180 (2016).
15. M. Tella-Amo et al., "A combined EM and visual tracking probabilistic model for robust mosaicking: application to fetoscopy," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit. Workshops*, pp. 84–92 (2016).
16. A. Heyden and M. Pollefeys, "Multiple view geometry," in *Emerging Topics in Computer Vision*, pp. 45–107, Cambridge University Press, Cambridge, England (2005).
17. S. J. Prince, Ed., *Computer Vision: Models, Learning, and Inference*, Cambridge University Press, Cambridge, England (2012).
18. D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vision* **60**(2), 91–110 (2004).
19. E. Rublee et al., "ORB: an efficient alternative to SIFT or SURF," in *Int. Conf. Comput. Vision*, IEEE, pp. 2564–2571 (2011).
20. H. Bay, T. Tuytelaars, and L. Van Gool, "SURF: speeded up robust features," in *Eur. Conf. Comput. Vision*, Springer, pp. 404–417 (2006).
21. R. Szeliski, "Image alignment and stitching: a tutorial," *Found. Trends® Comput. Graph. Vision* **2**(1), 1–104 (2006).
22. S. Benhimane and E. Malis, "Homography-based 2D visual tracking and servoing," *Int. J. Rob. Res.* **26**(7), 661–676 (2007).
23. M. Muja and D. G. Lowe, "Scalable nearest neighbor algorithms for high dimensional data," *IEEE Trans. Pattern Anal. Mach. Intell.* **36**, 2227–2240 (2014).
24. K. Pachtrachai et al., "Hand-eye calibration for robotic assisted minimally invasive surgery without a calibration object," in *IEEE/RSJ Intell. Rob. Syst. (IROS)*, IEEE, pp. 2485–2491 (2016).

Marcel Tella-Amo is a PhD student at the University College of London (UCL). He also received his MRes in medical imaging at UCL in 2016. Previously, he received his BE + MSc degree in telecom engineering focused on image processing at the Technical University of Catalonia (UPC), Barcelona. His research interests include machine learning and computer vision, specifically mosaicking techniques, and SLAM.

Loïc Peter is a postdoctoral research associate at the UCL. He received his PhD in computer science from the Technical University of Munich (TUM), Germany. Previous to that, he received a French engineering degree from the École Centrale Paris in 2011 and his MSc degree from the École Normale Supérieure de Cachan in applied mathematics.

Dzhoshkun I. Shakir is a research software engineer at the Wellcome/EPSRC Centre for Surgical and Interventional Sciences. He received his PhD from the Chair for Computer Aided Medical Procedures and Augmented Reality of the TUM. He received his MSc degree in computational science and engineering from TUM as well. His main research interest is intraoperative imaging. His work focuses on software and hardware engineering solutions for real-time medical imaging technologies.

Jan Deprest is an obstetrician-gynecologist with subspecialty in fetal medicine. He has been involved in instrument design for fetoscopic surgery, and his research is dedicated to experimental and clinical fetal therapies.

Danail Stoyanov is an associate professor at UCL. He received his PhD in computer science from Imperial College London specializing in medical image computing. Previous to that, he studied electronics and computer systems engineering at King's College London. His research interests and expertise are in surgical vision and computational imaging, surgical robotics, image-guided therapies, and surgical process analysis.

Tom Vercauteren is a professor of interventional image computing at King's College London where he holds the Medtronic/RAEng Research Chair in Machine Learning for Computer-Assisted Neurosurgery. Prior to this, he was an associate professor at UCL and deputy director of the Wellcome/EPSRC Centre for Interventional and Surgical Sciences (WEISS). He has ten years of MedTech industry experience with a first-hand track record in translating innovative research. His main interest is in translational computer-assisted intervention.

Sebastien Ourselin is head of the School of Biomedical Engineering and Imaging Sciences and chair of Healthcare Engineering at King's College London. He was a professor at UCL, London, United Kingdom. Prior to that, he founded and led the CSIRO BioMediA Lab, Australia. He is an associate editor for *Transactions on Medical Imaging*, the *Journal of Medical Imaging*, *Scientific Reports*, and *Medical Image Analysis*. His research interests include image registration, segmentation, statistical shape modeling, surgical simulation, image-guided therapy, and minimally invasive surgery.