

Probabilistic visual and electromagnetic data fusion for robust drift-free sequential mosaicking. Application to fetoscopy.

Marcel Tella-Amo^{a, *}, Loic Peter^a, Dzhoshkun I. Shakir^a, Jan Deprest^{a, c}, Danail Stoyanov^a, Juan Eugenio Iglesias^b, Tom Vercauteren^{a, c}, Sebastien Ourselin^a

^a Wellcome / EPSRC Centre for Interventional and Surgical Sciences, University College London, 66-72 Gower Street, London, UK, WC1E 6AA

^b University College London, Translational Imaging Group, CMIC, Medical Physics, 8.18, Malet Place Engineering Building, London, UK, WC1E 6B

^c KU Leuven, Centre for Surgical Technologies, Faculty of Medicine, Herestraat 49, Leuven, Belgium, 3000

Abstract. The most effective treatment for Twin-to-Twin Transfusion Syndrome is laser photocoagulation of the shared vascular anastomoses in the placenta. Vascular connections are extremely challenging to locate due to their caliber and the reduced field of view of the fetoscope. Therefore, mosaicking techniques are beneficial to expand the scene, facilitate navigation and allow vessel photocoagulation decision-making. Local vision-based mosaicking algorithms inherently drift over time due to the use of pairwise transformations. We propose the use of an electromagnetic tracker (EMT) sensor mounted at the tip of the fetoscope to obtain camera pose measurements, which we incorporate into a probabilistic framework with frame-to-frame visual information to achieve globally consistent sequential mosaics. We parametrize the problem in terms of plane and camera poses constrained by EMT measurements to enforce global consistency while leveraging pairwise image relationships in a sequential fashion through the use of Local Bundle Adjustment. We show that our approach is drift-free and performs similarly to state-of-the-art global alignment techniques like Bundle Adjustment albeit with much less computational burden. Additionally, we propose a version of Bundle Adjustment that uses EMT information. We demonstrate the robustness to EMT noise and loss of visual information and evaluate mosaics for synthetic and phantom-based datasets.

Keywords: Fetoscopy, Laser Coagulation, TTTS, sequential mosaicking, globally consistent mosaicking, electromagnetic tracking, sensor fusion .

* Marcel Tella-Amo, marcel.tella.14@ucl.ac.uk

1 Introduction

Twin-to-Twin Transfusion Syndrome (TTTS) complicates 10-15% of monochorionic diamniotic (MCDA) pregnancies.¹ Monochorionic twins share a single placenta and their circulation due to the presence of inter-twin anastomoses. A certain unfavourable pattern may result in an imbalance of inter-twin blood flow, leading to acute, mid-trimester TTTS. It causes overproduction of urine (hence polyhydramnios) in the recipient, while the other fetus will have oligohydramnios. Due to the acute overdistention of the uterus, mothers may go into labour or rupture their membranes. TTTS can also lead to cardiac dysfunction in one or both fetuses, worsening the prognosis. If this

condition is not treated, the outcome is nearly always fatal.²

The standard of care today is fetoscopic laser photocoagulation,³ which has been shown to be more effective than serial removal of excessive amniotic fluid.⁴ The procedure consists of the insertion of a fetoscope, identification and coagulation of all visible anastomoses, and functionally disconnect the two circulations.

The success of this operation is dependent on many variables, some of them related to the operative technique. The surgeon has to be able to inspect as much of the placenta and understand its angioarchitecture. Ideally one obtains a general insight on the nature (whether the connecting vessels are arteries or veins) and location of all anastomoses. Following this, first arterio-venous and then veno-arterial connections are coagulated. At the end of the procedure, it is recommended to superficially laser the area between the lasered anastomoses to avoid the persistence of flow over non-visualized, smaller anastomoses (referred to as the Solomon or bichorionization technique²). Fetoscopy is typically performed with 1.3 to 2.0 mm fiberendoscope and limited light. The most limiting factor for keeping an overview of the vascular anatomy is the small field of view.

To address this limitation, the creation of a 2D mosaic of the placenta has been proposed⁵⁻⁷ as a means of expanding the field of view by stitching the fetoscopic images to a common reference frame.

The number of images needed to cover the whole placenta are an additional challenge. Clinical imaging conditions are also restrictive due to a lack of visual texture and color contrast between arteries and veins, in particular of small diameter, and visual artifacts such as blood, amniotic fluid particles, the presence of the inter-twin membrane or even fetal movements, which may perturb the vision. To leverage the use of imagery, there should be a large overlap between adjacent images to ensure a successful registration at the expense of increasing, even more, the number of frames.

Local vision-based mosaicking algorithms make use of pairwise transformations between images to compose a mosaic. This has a fundamental limitation; since the transformation of each frame to the reference space is made dependent of all the previous pairwise registrations, any new pairwise registration error is propagated through a chain of transformations. As a result, an inevitable, progressive drift in the reference space occurs, which can even degenerate in a rupture of the chain of transformation if a pair of images cannot be registered. In order to illustrate this effect on a simulated dataset, Fig. 1a shows a ground truth mosaic composed of 200 images where the camera has moved following a circular pattern. Figure 1b shows a mosaic that has experienced drift due to the composition of homographies.

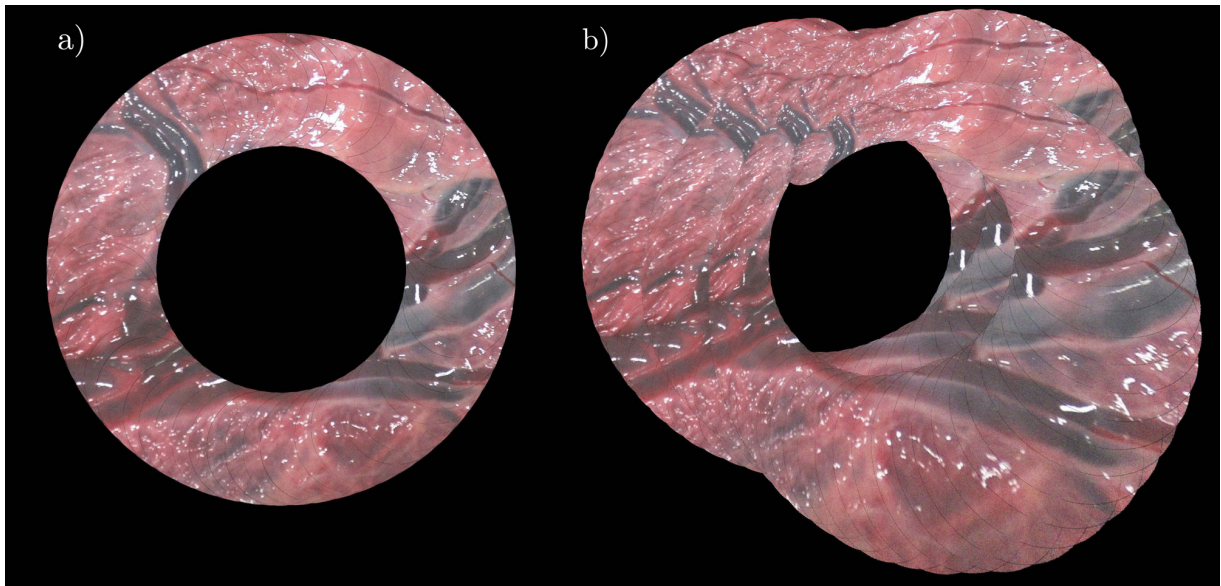


Fig. 1 Mosaics of 200 synthetic images where the camera has moved following a circular pattern of four laps. a) The ground truth mosaic. b) A mosaic with drift due to the accumulation of error in subsequent iterations.

To address this problem, we propose the use of an Electromagnetic Tracking system (EMT) by attaching an EMT sensor⁸⁻¹³ to the tip of the fetoscope. The EMT system provides measurements of the 3D pose of the sensor, which then relate to the 3D pose of the fetoscope through a pre-computed rigid Hand-eye calibration matrix. These measurements do not provide enough

information to create a mosaic since the geometry of the scene is unknown. In fact, even if the geometry of the scene was known, the noise or jitter in these measurements would propagate to the mosaic space, resulting in misregistrations in the mosaic. Hence, the generation of the mosaic using exclusively the EMT information is infeasible (See Sec. 4.2.2). However, the fusion of these measurements is extremely valuable in order to guide the estimation of the mosaic and prevent it from drifting; especially in fetoscopy, where the poor quality of the pairwise registrations can accentuate the drift.

In this paper, we present a probabilistic model that uses the complementarity between the EMT and visual information to drive the estimation towards globally consistent mosaics i.e. that do not suffer from drift, independently of the number of frames. Additionally, we compare our algorithm with the state-of-the-art in global alignment^{14,15} i.e. Bundle Adjustment, and show that we achieve a similar performance to the state-of-the-art with a much lower computational burden.

The improvement in terms of computational complexity is mostly observed in two main steps: the matching and the non-linear optimization. Since in Bundle Adjustment, the number of image pairs where matching needs to be attempted is N^2 in the worst case, where N is the number of images in the sequence, this stage of the algorithm has a complexity of $\mathcal{O}(N^2)$. In the proposed algorithm, only the images within a fixed window are considered for matching. Therefore, only a fixed number of images proportional to N is matched ($\mathcal{O}(N)$). Analogously, the visual cost function computes the residuals for all the obtained correspondences in potentially N^2 pairs of images ($\mathcal{O}(N^2)$) and estimates all the parameters at a time while our algorithm computes only the visual residual within the window ($\mathcal{O}(N)$) optimizing only for a small subset of parameters.

Finally, we show that our algorithm is robust to EMT noise and loss of visual information and evaluate mosaics for synthetic and phantom-based datasets. Additionally, we propose a version of

Bundle Adjustment that incorporates EMT information in an analogous manner.

The rest of the paper is structured as follows. In Sec. 2, we discuss the state-of-the-art in drift-free mosaicking using visual information as well as sensor fusion. Section 3 details our approach and Sec. 4 provides an experimental evaluation. Finally, we discuss the results, the limitations of the approach, and propose future research lines to overcome them.

2 Related Work

Mosaicking algorithms have been extensively explored in computer vision for the last two decades. Vision-based mosaicking generally relies on estimating transformations with respect to the mosaic space by chaining transformations between adjacent images.^{14–16} Therefore, it inherently accumulates drift. Additionally, if one of the transformations fails to be estimated, the resulting mosaic cannot be computed further.

Michaelsen¹⁷ observed that a translation-based mosaic accumulates less drift since no multiplicative effect applies. He presented a patch-wise algorithm where the normal vector to the imaged patch is used to correct the homography between the first and last image in the patch such that only x-y translation components are used to stitch the patch into the global reference frame.

As an alternative strategy, Sawhney et al.¹⁸ explored the idea of iteratively finding the topology of the images. First using a translation-based registration, the authors check for overlapping images to determine the neighborhood of a given image. Then, they use a projective mapping to estimate locally consistent patches within the neighborhood. Finally, a global refinement of the mosaic is performed to avoid local minima due to the local refinement of the patches. Given that the EMT system provides the position of the cameras, we directly have information about the topology and we aim for a sequential estimate that is globally consistent.

Given that the accumulation of error leads to non-matching image positions when revisiting the same scene, so-called loop-closing strategies propose to identify and add the loop closure overlapping frames as an additional constraint¹⁹⁻²¹ and then correct for the accumulation of error in the rest of the loop. Civera et al.²² keep track of the observed features and optimize for all the camera poses in every iteration, having a natural loop closing effect since the whole loop is optimized for. This has the additional cost of having to compare all images to a growing map in each iteration. The creation of this map is not trivial in our scenario given the poor fetoscopic image quality: the lack of visual texture and color contrast between arteries and veins, amniotic fluid particles, the presence of the inter-twin membrane and fetal movements.

Another cause of this accumulation of error in a planar scenario is over-parametrization.²³ In a planar scenario, the family of homographies that defines the motion of a monocular camera can be minimally parametrized by six parameters for each camera pose and the three global parameters representing the plane whereas in classic mosaicking every pairwise relation is parametrized by a full homography. There exists then a restricted group of homographies that can each be decomposed into the camera motion and global plane. This decomposition has been used by several authors, for example, Benhimane and Malis²⁴ used it for visual servoing and some authors used it for mosaicking as well.^{7,17,23,25}

Closely related to the last idea, Olsson and Eriksson²⁶ showed an increased performance in the estimation of the plane by minimizing the reprojection errors as a function of the plane as opposed to using a plane fitting procedure after triangulating the 3D points. Given that depth is an unobserved quantity, if there is not enough observability and baseline, depth estimates of the 3D points may not be accurate and therefore fitting a plane leads to poorer performance than estimating an underlying plane modeling the structure. While they assume that the camera positions are

provided, we estimate them from visual information and noisy EMT measurements.

The state-of-the-art in terms of drift-free alignment is the well-known Bundle Adjustment.^{14,15} This is a batch non-linear optimization that minimizes the reprojection residuals in all images. As an example close to our application, Atasoy et al.²⁷ proposed a vision-based version of Bundle Adjustment for fibroscopic video mosaicking that weights the images with the number of matched features found in each pair. In Sec. 3, we introduce a problem-specific version of Bundle Adjustment that fuses the visual and EMT data.

Mur-Artal et al.¹⁹ showed that it is possible to meet real-time performance requirements with the use of a windowed iterative approach also known as Local Bundle Adjustment. Under the assumption that the global minimum is too expensive to reach and that the information is provided in an incremental fashion, this algorithm computes the estimates in a window with the underlying idea that cameras outside the window provide little information about the current estimates. Therefore, the estimated cameras are considered fixed in the following iteration and only new cameras are to be estimated. This yields accurate estimates, usually close to the Bundle Adjustment.²⁸ Consequently, we use it in our application to achieve sequential yet accurate mosaics.

The use of an external sensor in our framework aims to constrain the global pose estimates in order to bound the drift. Similar to ours is the work of Agrawal et al.,²⁹ who explored the idea of using an inexpensive Global Positioning System (GPS) combined with a stereo vision system. They used a Kalman filter limiting the drift in translation for long robot trajectories. In our scenario, the fetoscope is close to the tissue and its movement follows a hand-held pattern, which implies additional complications.

Vyas et al.¹² also integrated an EMT system in a mosaicking pipeline. In their system, the camera movement is restricted to fronto-parallel motion. The registration consists of two steps:

first, the images are placed according to the EMT measurements and then, a pairwise adjustment is performed using a cross-correlation. This method does not make optimal use of the available information since it only enforces pairwise consistency and the optimization does not take into account the electromagnetic data. In contrast, our probabilistic integration leverages the fact that the EMT measurements are centered in the true camera position and uses jointly either all or a larger subset of information available.

There has been extensive work on the fusion of other types of sensors. The integration of gyroscopes,³⁰ accelerometers and Inertial Measurement Units (IMU)³¹ with visual data has been shown to achieve better accuracy and robustness in the estimation of homographies. Additionally, in,³¹ an IMU combined with visual keypoints are integrated in a fully probabilistic manner in order to improve the robustness and accuracy of the estimation in a SLAM system. Another idea exploited by some authors is to use predictions of the camera poses and the inertial measurements to re-calibrate the bias term^{32–34} inherent in inertial measurement systems in order to eliminate the drift. To our knowledge, it is not currently possible to fit an inertial system into a clinical fetoscope due to its dimensions.

In our previous work,⁷ we proposed a preliminary model to reduce the drift using the EMT and visual information jointly. In the current work, we extend our framework and validate the complete elimination of the drift.

3 Methods

First, we present preliminaries in mosaicking in Sec. 3.1. Then, Sec. 3.2 introduces the EMT measurements and the relationships used to incorporate them into the mosaicking pipeline. In Sec. 3.3, we detail our main contribution; the formulation of a probabilistic model that achieves

a sequential drift-free estimate of the mosaic independently of the number of frames using the complementarity between visual and EMT information. We then state the assumptions of the model and explain the parametrization used. Finally, we introduce the two proposed algorithms to do inference in Sec. 3.4.

3.1 Preliminaries in Mosaicking

Given a sequence of 2D images $\mathcal{I} = \{\mathbf{I}_k\}_{k=1}^N$ of a planar scene acquired by a hand-held monocular camera with limited field-of-view, where $\pi \in \mathbb{R}^3$ denotes the plane, we seek to find a representation of the scene, the mosaic $\mathbf{M} : \Omega_M \rightarrow \mathbb{R}^3$ where $\Omega_M \subset \mathbb{R}^2$, that captures the entire observed area into a two-dimensional space where each coordinate is associated with the RGB components of a pixel in the mosaic.

Provided that the structure is a plane, a homography $h(\cdot) : \mathbb{R}^2 \rightarrow \mathbb{R}^2$, which can also be expressed as a matrix $\mathbf{H} \in \mathbb{R}^{3 \times 3}$ in homogeneous coordinates, maps the location of any point $\mathbf{p} = \begin{bmatrix} p_x & p_y \end{bmatrix}^T$ in an image to its corresponding point $\mathbf{p}' = \begin{bmatrix} p'_x & p'_y \end{bmatrix}^T$ in a second image

$$\begin{aligned} p'_x &= \frac{H_{1,1}p_x + H_{1,2}p_y + H_{1,3}}{H_{3,1}p_x + H_{3,2}p_y + H_{3,3}} \\ p'_y &= \frac{H_{2,1}p_x + H_{2,2}p_y + H_{2,3}}{H_{3,1}p_x + H_{3,2}p_y + H_{3,3}}. \end{aligned} \quad (1)$$

If expressed as a matrix, Eq. (1) can be written as $\tilde{\mathbf{p}}' \propto \mathbf{H}\tilde{\mathbf{p}}$ where \mathbf{H} is known up to a scale factor.³⁵ We use a tilde in an image point to indicate homogeneous coordinates.

We can relate image k to image j of the sequence with a chain of homographies as follows:

$$\mathbf{H}_{j,k} = \prod_{l=k}^{j-1} \mathbf{H}_{l+1,l} \quad (2)$$

where the product operator denotes the left matrix multiplication. To build a mosaic, we need to define the common space Ω_M where all images are stitched. Without loss of generality, if we choose the space of the first image as the mosaic space, then Eq. (2) with $k = 1$ expresses the relation between any image j in the sequence and the mosaic space. Throughout this manuscript, a homography with only one sub-index relates an image j to the mosaic space e.g. $\mathbf{H}_j = \mathbf{H}_{j,1}$ whereas a homography with two sub-indexes relates two images e.g. $\mathbf{H}_{j,k} = \mathbf{H}_j \mathbf{H}_k^{-1}$ relates image k to image j .

A pairwise homography $\mathbf{H}_{j,k}$ can be directly obtained if a part of the scene is present in both images. In this work, we use a landmark-based approach^{36–38} to find correspondences in the images. Once the correspondences are computed, an approximation of the homography can be estimated using the DLT algorithm and further refined through a non-linear optimization.³⁹ However, the estimation of a homography between two images inevitably carries error, which leads to accumulation of error when propagated through the chain in Eq. (2). We propose to tackle this problem by incorporating measurements given from an EMT system.

3.2 Incorporation of the EMT System

We propose to bound the drift in the mosaic by relying on a set of camera pose measurements $\mathcal{Z} = \{\mathbf{z}_k\}_{k=1}^N$ provided by the EMT system. In order to use EMT measurements in conjunction with visual information, we must establish a link between them. To this end, we place a virtual camera at the origin of an arbitrary global coordinate system, whose image plane can be set to coincide with the mosaic space Ω_M . Then, the image plane from the camera pose at time k is related to the image plane of the virtual camera by its homography \mathbf{H}_k as illustrated in Fig. 2.

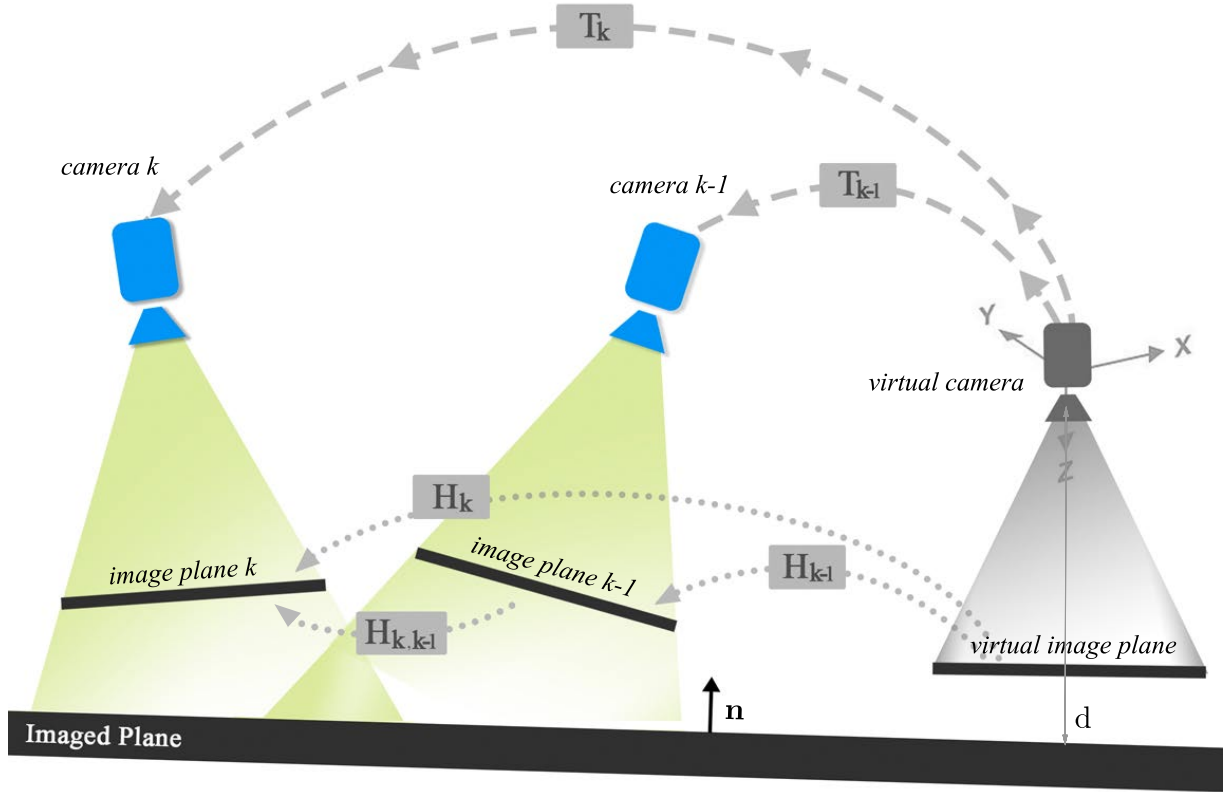


Fig. 2 A virtual camera is placed at the origin of an arbitrary global coordinate system. The relation between images and cameras is defined in Eqs. (3) and (5).

Therefore, we can establish the relation

$$\mathbf{H}_{k,k-1} = \mathbf{H}_k \mathbf{H}_{k-1}^{-1}. \quad (3)$$

Moreover, a 3D point \mathbf{p}_{3D} on the imaged plane satisfies $\begin{bmatrix} \mathbf{n}^T & d \end{bmatrix} \begin{bmatrix} \mathbf{p}_{3D} \\ 1 \end{bmatrix} = 0$ where the unit

vector $\mathbf{n} = \begin{bmatrix} n_x & n_y & n_z \end{bmatrix}^T$ and the distance d from the virtual camera to the plane are seen from the point of view of the virtual camera. Let $\mathbf{T} \in \mathbb{S}\mathbb{E}(3)$ be a camera pose expressed as rigid body

transformation in 3D space

$$\mathbf{T} = \begin{bmatrix} \mathbf{R} & \mathbf{t} \\ \mathbf{0} & 1 \end{bmatrix} \quad (4)$$

where \mathbf{R} is a rotation matrix in $\mathbb{SO}(3)$ and \mathbf{t} is a translation vector in \mathbb{R}^3 . Provided that we want to incorporate camera pose measurements, we establish the link between the true camera poses $(\mathbf{T}_k, \mathbf{T}_{k-1})$ that induce the homographies $(\mathbf{H}_k, \mathbf{H}_{k-1})$ through the plane^{24,35} as

$$\mathbf{H}_k = \mathbf{K}(\mathbf{R}_k - \mathbf{t}_k \frac{\mathbf{n}^T}{d})\mathbf{K}^{-1} \quad (5)$$

where \mathbf{K} is the pre-calibrated intrinsic camera matrix.

If the plane π was known, then we could compose a mosaic only using the EMT measurements and the images. However, the EMT noise in the camera pose measurements propagates to the mosaic space causing a jitter effect that translates into misregistrations in the composition. More importantly, we do not know π a priori. Therefore, while the guidance of the EMT system is of crucial importance for global positioning consistency, it is necessary to combine both EMT and visual information to integrate knowledge of the scene, implicitly estimating the plane, and obtain pairwise registrations that are as accurate as possible. To this end, we propose a generative probabilistic model that seeks the set of camera poses $\mathcal{X} = \{\mathbf{x}_k\}_{k=1}^N$ and plane π that generated the sequence of images \mathcal{I} and the EMT measurements \mathcal{Z} , which are then used to project the images to the mosaic space Ω_M and create the mosaic \mathbf{M}

$$(\hat{\mathcal{X}}, \hat{\pi}) = \underset{(\mathcal{X}, \pi)}{\operatorname{argmax}} P(\mathcal{X}, \pi | \mathcal{Z}, \mathcal{I}). \quad (6)$$

We now detail the notation, assumptions, and parametrization used in the probabilistic model.

3.3 The Probabilistic Model

3.3.1 Notation and modeling assumptions

Consider two images. Let image A be the source image and image B be the target image. These images each contain a set $\mathcal{P}_{l,m}^A$ and $\mathcal{P}_{l,m}^B$ of $N_{l,m}$ corresponding landmarks found from image m to image l . For each set, let us define the i th corresponding landmark in the set as $\{\mathbf{p}_{l,m}^{A,i} \in \mathbb{R}^2 | \mathcal{P}_{l,m}^A = \{\mathbf{p}_{l,m}^{A,i}\}_{i=1}^{N_{l,m}}\}$ for image A and $\{\mathbf{p}_{l,m}^{B,i} \in \mathbb{R}^2 | \mathcal{P}_{l,m}^B = \{\mathbf{p}_{l,m}^{B,i}\}_{i=1}^{N_{l,m}}\}$ for image B. For simplicity, we assume that landmarks in an image are independent. Additionally, for different pairs of images, we consider independence of all the sets in source images in $\mathcal{P}^A = \bigcup_{l,m \in \mathcal{L}} \mathcal{P}_{l,m}^A$, and target images $\mathcal{P}^B = \bigcup_{l,m \in \mathcal{L}} \mathcal{P}_{l,m}^B$, where \mathcal{L} is the set of all possible corresponding image indexes. Figure 3 depicts a schematic of the nomenclature of the correspondences.

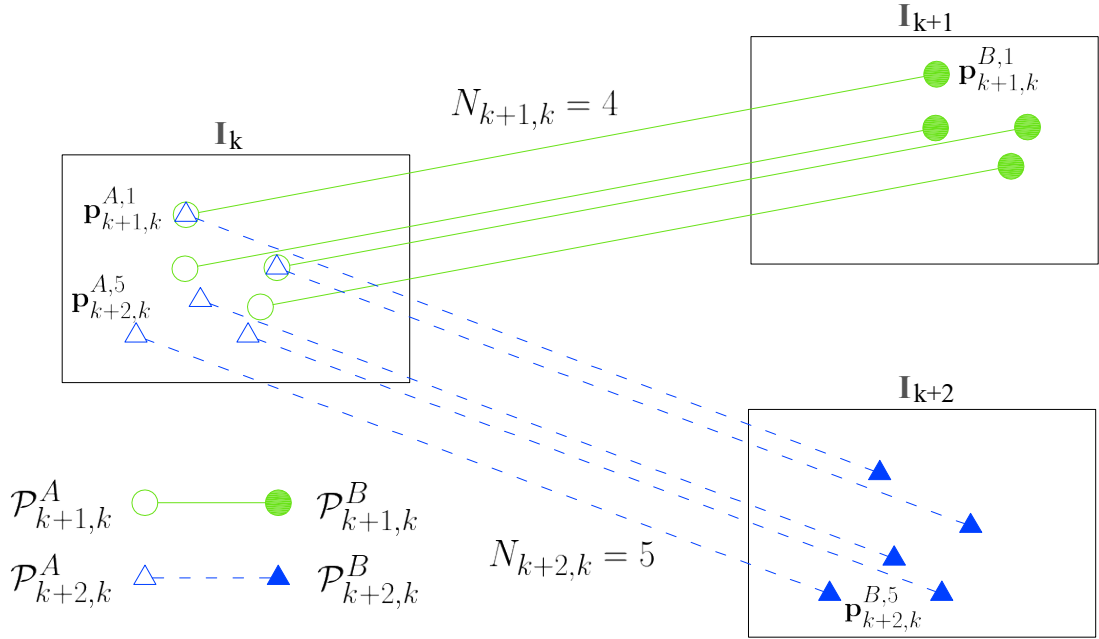


Fig. 3 Schematic of the nomenclature of the correspondences. $\mathcal{P}_{k+1,k}^A$ (green empty circles) is the set of points of image A for which $N_{k+1,k} = 4$ correspondences from image k to $k + 1$ are available, being $\mathcal{P}_{k+1,k}^B$ (green colored circles) the corresponding landmarks in image B. $\mathcal{P}_{k+2,k}^A$ (blue empty triangles) is the set of points of image A for which $N_{k+2,k} = 5$ correspondences from image k to $k + 2$ are available, being $\mathcal{P}_{k+2,k}^B$ (blue colored triangles) the corresponding landmarks in image B. As an example, the landmark $\mathbf{p}_{k+1,k}^{A,1}$ corresponds to the point $\mathbf{p}_{k+1,k}^{B,1}$ and the point $\mathbf{p}_{k+2,k}^{A,5}$ corresponds to the point $\mathbf{p}_{k+2,k}^{B,5}$.

In terms of the parametrization, we use a scaled normal vector $\pi = \mathbf{n}/d \in \mathbb{R}^3$ to parametrize the plane. Compared to other parametrizations,²³ this has the advantage of encoding the inverse depth ($1/d$) in each of the components, reducing the non-linearity and thus accelerating the convergence. We use a minimal parametrization of six parameters for the camera poses $\mathbf{x}_k = \left[\mathbf{r}_k^T \quad \mathbf{t}_k^T \right]^T$ with the orientation $\mathbf{r} = \begin{bmatrix} r_x & r_y & r_z \end{bmatrix}^T$ being the Euclidean vector in \mathbb{R}^3 identified with the skew-symmetric matrix $\mathbf{S} \in \mathfrak{so}(3)$ for which $\mathbf{R} = e^{\mathbf{S}} \in \mathbb{SO}(3)$

$$\mathbf{S} = \begin{bmatrix} 0 & -r_z & r_y \\ r_z & 0 & -r_x \\ -r_y & r_x & 0 \end{bmatrix}. \quad (7)$$

This parametrization ensures valid rotation matrices and is valid for all angles as the exponential map in $\mathfrak{so}(3)$ is surjective. Furthermore, the camera is guaranteed to look downwards, towards the placenta, and therefore the exponential map is bijective for all angles of interest as long as the camera does not complete a full rotation around its z axis, which is very unlikely.

Once the notation has been defined, we state the main modeling assumptions:

1. We consider the imaged object to be a plane.⁷
2. Every EMT measurement \mathbf{z}_k is modeled as a Gaussian random variable centered on the true camera pose \mathbf{x}_k with diagonal covariance Σ_{EMT} , that is $\mathbf{z}_k | \mathbf{x}_k \sim \mathcal{N}_{\mathbf{z}_k}(\mathbf{x}_k, \Sigma_{EMT})$. Even though the EMT measurements are actually not strictly Gaussian,^{40,41} this is a common assumption⁴⁰ that simplifies the problem. We account for modeling errors in the EMT measurements by enlarging its standard deviation.
3. The locations of corresponding points between adjacent images match the same visual con-

tent but they are imperfect; each pair of corresponding points has a matching error defined as the distance between a point in an image and its correspondence propagated from the other image. For simplicity, we assume this error to have zero mean and diagonal covariance matrix $\sigma_v^2 \mathbf{I}$. This is equivalent to saying that a 2D point in image B is a Gaussian measurement generated from a true 2D point in image A, that is $\mathbf{p}_{l,m}^{B,i} | \mathbf{x}_l, \mathbf{x}_m, \pi, \mathbf{p}_{l,m}^{A,i} \sim \mathcal{N}_{\mathbf{p}_{l,m}^{B,i}}(\mu_v^i, \sigma_v^2 \mathbf{I})$ in which the mean $\mu_v^i(\mathbf{x}_l, \mathbf{x}_m, \pi, \mathbf{p}_{l,m}^{A,i})$, detailed later in Eq. (13), is the projected location of the point in image A. See the discussion section for more comments about this modeling assumption.

4. We assume that the camera is moving smoothly and therefore model the relation between camera poses with a constant velocity motion model⁴² with mean $\mu_p(\mathbf{x}_{k-1}, \mathbf{x}_{k-2})$, which is described later in Eq. (15), and diagonal covariance matrix Σ_p , that is $\mathbf{x}_k | \mathbf{x}_{k-1}, \mathbf{x}_{k-2} \sim \mathcal{N}_{\mathbf{x}_k}(\mu_p, \Sigma_p)$. This model expresses that the velocity at time k must be the same as the velocity at time $k - 1$ plus a perturbation.

Within this probabilistic framework, the estimation of the mosaic can be cast as a Bayesian inference problem, in which the posterior $P(\mathcal{X}, \pi | \mathcal{Z}, \mathcal{P}^A, \mathcal{P}^B)$ is maximized with respect to the camera poses \mathcal{X} and plane π

$$(\hat{\mathcal{X}}, \hat{\pi}) = \underset{(\mathcal{X}, \pi)}{\operatorname{argmax}} P(\mathcal{X}, \pi | \mathcal{Z}, \mathcal{P}^A, \mathcal{P}^B) \quad (8)$$

in which the posterior probability factorizes as

$$\begin{aligned} P(\mathcal{X}, \pi | \mathcal{Z}, \mathcal{P}^A, \mathcal{P}^B) &\propto P(\mathcal{Z}, \mathcal{P}^B | \mathcal{X}, \pi, \mathcal{P}^A) P(\mathcal{X}, \pi | \mathcal{P}^A) \\ &\propto \underbrace{P(\mathcal{Z} | \mathcal{X})}_{(a)} \underbrace{P(\mathcal{P}^B | \mathcal{X}, \pi, \mathcal{P}^A)}_{(b)} \underbrace{P(\mathcal{X}) P(\pi)}_{(c)}. \end{aligned} \quad (9)$$

First, we applied Bayes Theorem and dropped the constant factor. Second, if the true cameras \mathcal{X} are given, then the EMT measurements \mathcal{Z} are independent of the visual terms \mathcal{P}^B and therefore, $P(\mathcal{Z}, \mathcal{P}^B | \mathcal{X}, \pi, \mathcal{P}^A)$ can be separated as (a) and (b). Furthermore, in (a), we applied the fact that the EMT measurements are independent of the plane π given \mathcal{X} so that $P(\mathcal{Z} | \mathcal{X}, \pi) = P(\mathcal{Z} | \mathcal{X})$.

In the prior term identified as (c), we have first considered independence of \mathcal{P}^A and π and then independence of π and \mathcal{X} , i.e. $P(\mathcal{X}, \pi | \mathcal{P}^A) = P(\mathcal{X}, \pi) = P(\mathcal{X})P(\pi)$. We assume that we do not have prior information about the plane and that its distribution is bounded, considering thus $P(\pi) \propto 1$. The graphical model of the proposed probabilistic framework is presented in Fig. 4. Circles represent random variables, which can be either latent (white background) or observed (shaded). Parameters of the model are depicted as a point.

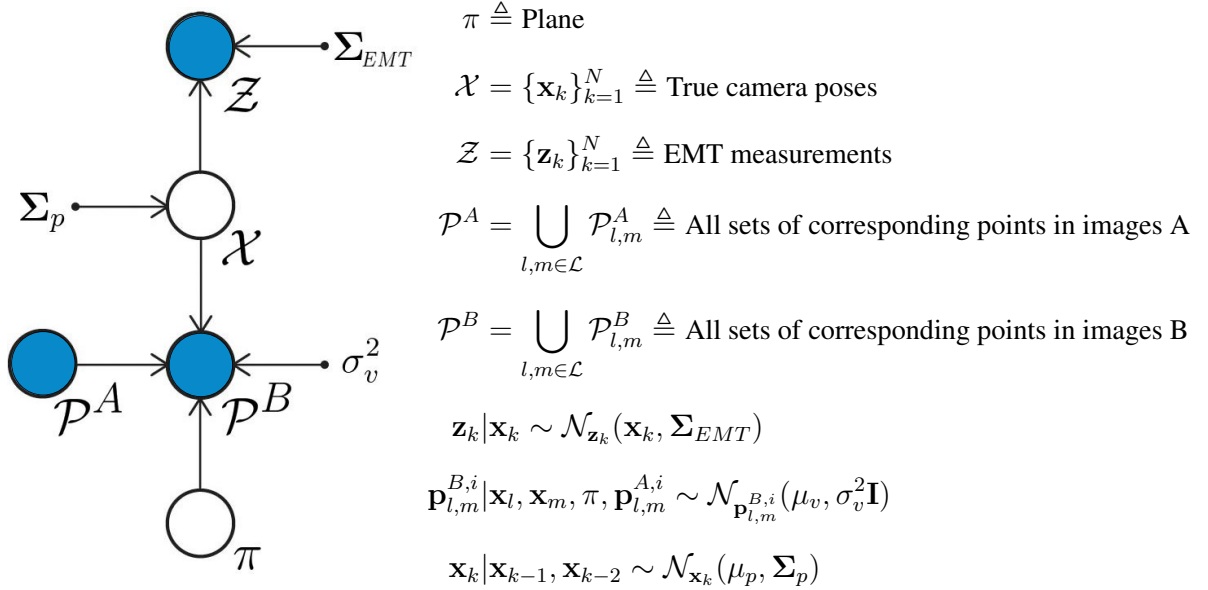


Fig. 4 Graphical model of the proposed probabilistic framework. Circles represent random variables, which can be either latent (white background) or observed (shaded). Parameters of the model are depicted as a point.

Next, we incorporate the assumptions and present the factorization of the likelihood and prior terms.

3.3.2 The likelihood and prior

The likelihood combines the EMT term $P(\mathcal{Z}|\mathcal{X})$ and visual term $P(\mathcal{P}^B|\mathcal{X}, \pi, \mathcal{P}^A)$. In particular, the EMT term contains the relation between the EMT measurements and the true cameras, which assuming independence of the EMT measurements, can be expressed as

$$P(\mathcal{Z}|\mathcal{X}) = \prod_{k=1}^N P(\mathbf{z}_k|\mathbf{x}_k). \quad (10)$$

The visual term comes from the correspondences between images and it enforces visual pairwise consistency. Applying the assumption that sets of corresponding points are independent of each other, we can simplify this term by factorizing it as

$$P(\mathcal{P}^B|\mathcal{X}, \pi, \mathcal{P}^A) = \prod_{\{m,l\} \in \mathcal{L}} P(\mathcal{P}_{l,m}^B|\mathbf{x}_l, \mathbf{x}_m, \pi, \mathcal{P}_{l,m}^A) \quad (11)$$

In every image, we also assume every landmark to be independent. Therefore, we can further decouple the points as

$$P(\mathcal{P}_{l,m}^B|\mathbf{x}_l, \mathbf{x}_m, \pi, \mathcal{P}_{l,m}^A) = \prod_{i=1}^{M^{l,m}} P(\mathbf{p}_{l,m}^{B,i}|\mathbf{x}_l, \mathbf{x}_m, \pi, \mathbf{p}_{l,m}^{A,i}). \quad (12)$$

We assume every point in image B to be $\mathbf{p}_{l,m}^{B,i}|\mathbf{x}_l, \mathbf{x}_m, \pi, \mathbf{p}_{l,m}^{A,i} \sim \mathcal{N}_{\mathbf{p}_{l,m}^{B,i}}(\mu_v^i, \sigma_v^2 \mathbf{I})$ with

$$\begin{bmatrix} \mu_v^i(\mathbf{x}_l, \mathbf{x}_m, \pi, \mathbf{p}_{l,m}^{A,i}) \\ 1 \end{bmatrix} \propto \mathbf{H}_l \mathbf{H}_m^{-1} \tilde{\mathbf{p}}_{l,m}^{A,i} = (\mathbf{R}_l - \mathbf{t}_l \frac{\mathbf{n}^T}{d})(\mathbf{R}_m - \mathbf{t}_m \frac{\mathbf{n}^T}{d})^{-1} \tilde{\mathbf{p}}_{l,m}^{A,i} \quad (13)$$

For convenience, we redefine every point $\tilde{\mathbf{p}}$ directly in the normalized image space through $\tilde{\mathbf{p}} = \mathbf{K}^{-1}\tilde{\mathbf{q}}$ being $\tilde{\mathbf{q}}$ a homogeneous point in the image space.

The number of correspondences has a strong impact on the estimation. A high number emphasizes the visual term, unbalancing the fusion with the EMT information. This effect is related to the simplifying independence assumption between landmarks. The solution adopted has been to normalize and manually adapt the visual variance σ_v^2 to ensure good inter-frame registrations, which can be seen as a pragmatic correction factor. A more detailed explanation of this problem is covered in the discussion section.

The prior term $P(\mathcal{X})$ on the camera poses is approximated by using a second order Markov process which accounts for a constant velocity of the camera. We assume no prior knowledge of the joint probability on a bounded region of the space, i.e. $P(\mathbf{x}_2, \mathbf{x}_1) \propto 1$

$$\begin{aligned} P(\mathcal{X}) &= P(\mathbf{x}_1, \dots, \mathbf{x}_N) = P(\mathbf{x}_2, \mathbf{x}_1) \prod_{k=3}^N P(\mathbf{x}_k | \mathbf{x}_{k-1}, \mathbf{x}_{k-2}, \dots, \mathbf{x}_1) \\ &\approx P(\mathbf{x}_2, \mathbf{x}_1) \prod_{k=3}^N P(\mathbf{x}_k | \mathbf{x}_{k-1}, \mathbf{x}_{k-2}) \propto \prod_{k=3}^N P(\mathbf{x}_k | \mathbf{x}_{k-1}, \mathbf{x}_{k-2}). \end{aligned} \quad (14)$$

We assume the new camera motion $\mathbf{x}_k | \mathbf{x}_{k-1}, \mathbf{x}_{k-2} \sim \mathcal{N}_{\mathbf{x}_k}(\mu_p, \Sigma_p)$, where

$$\mu_p(\mathbf{x}_{k-1}, \mathbf{x}_{k-2}) = \begin{bmatrix} \bar{\mathbf{r}}_k^T & \bar{\mathbf{t}}_k^T \end{bmatrix}^T \quad (15)$$

is decomposed from the rigid transformation $\bar{\mathbf{T}}_k$ into rotation and translation

$$\bar{\mathbf{T}}_k = \begin{bmatrix} \bar{\mathbf{R}}_k & \bar{\mathbf{t}}_k \\ \mathbf{0} & 1 \end{bmatrix} = \mathbf{T}_{k-1, k-2} \mathbf{T}_{k-1} \quad (16)$$

with $\bar{\mathbf{R}}_k = \exp(\bar{\mathbf{r}}_k)$. This simply says that the last camera pose \mathbf{T}_{k-1} is composed with the last available pairwise velocity $\mathbf{T}_{k-1,k-2} = \mathbf{T}_{k-1} \mathbf{T}_{k-2}^{-1}$, giving an approximate idea of where the current estimate should be.

3.4 Inference

By applying a negative logarithm to the posterior probability distribution, we can express the proposed model as the minimization of a cost which contains three terms, the visual cost C_v , the EMT cost C_{EMT} and the cost associated with the temporal model C_p as

$$(\hat{\mathcal{X}}, \hat{\pi}) = \underset{(\mathcal{X}, \pi)}{\operatorname{argmin}} (C_v + C_{EMT} + C_p) \quad (17)$$

where

$$C_v = \sum_{l,m \in \mathcal{L}} \sum_{i=1}^{N_{l,m}} \frac{1}{\sigma_v^2} \|\mathbf{p}^{B,i} - \mu_v^i(\mathbf{x}_l, \mathbf{x}_m, \pi, \mathbf{p}_{l,m}^{A,i})\|_2^2 \quad (18)$$

$$C_{EMT} = \sum_{k=1}^N (\mathbf{z}_k - \mathbf{x}_k)^T \boldsymbol{\Sigma}_{EMT}^{-1} (\mathbf{z}_k - \mathbf{x}_k) \quad (19)$$

$$C_p = \sum_{k=3}^N (\mathbf{x}_k - \mu_p(\mathbf{x}_{k-1}, \mathbf{x}_{k-2}))^T \boldsymbol{\Sigma}_p^{-1} (\mathbf{x}_k - \mu_p(\mathbf{x}_{k-1}, \mathbf{x}_{k-2})). \quad (20)$$

This is a large scale non-linear least squares problem, which can be solved using a Gauss-Newton method, for which the EMT measurements can be used for initialization.

If only the visual cost is used, the problem results in Bundle Adjustment. Therefore, the proposed algorithm is an adapted version of Bundle Adjustment that also incorporates the EMT measurements and temporal consistency of the camera motions. These algorithms require to have all the information beforehand, i.e. they are offline. This is prohibitive in our case since we aim for

a sequential estimate. Consequently, we move towards local methods. However, we do consider the incorporation of the EMT information in Bundle Adjustment and propose it as an additional contribution, since it can be used for refinement at the end of the scanning procedure and can serve as a reference.

We use Local Bundle Adjustment (LBA), an approximation of Bundle Adjustment in which only the components within a temporal window of size W are considered. This reduces drastically the computational burden of the algorithm and allows for sequential estimation. We slightly modify this approach following these two main assumptions: (i) The cameras far from the current window provide little information about the new cameras to be estimated, yet they provide information about the plane given visual measurements. (ii) The cameras that have already been estimated are considered fixed in the next iteration. These assumptions are further commented in the discussion section.

Let $\mathcal{X}_e = \{\mathbf{x}_k^e\}_{k=1}^{\chi_e}$ be the subset of cameras to be estimated where χ_e is the number of cameras in the subset, $\mathcal{X}_g = \{\mathbf{x}_k^g\}_{k=1}^{\chi_g}$ the set of cameras already estimated and fixed within the window, with χ_g being the number of cameras in the subset as well as the index of the most recent fixed camera such that $W = \chi_e + \chi_g$. Additionally, let \mathcal{X}_o be the subset of cameras already estimated outside the window such that $\mathcal{X} = \{\mathcal{X}_e, \mathcal{X}_g, \mathcal{X}_o\}$. Analogously, let \mathcal{Z}_e be the set of EMT measurements corresponding to the camera poses to be estimated.

We now seek to maximize the posterior probability of the new camera motions \mathcal{X}_e and plane π given all the estimated cameras that have been fixed, and the EMT and visual measurements such that

$$(\hat{\mathcal{X}}_e, \hat{\pi}) = \underset{(\mathcal{X}_e, \pi)}{\operatorname{argmax}} P(\mathcal{X}_e, \pi | \mathcal{X}_g, \mathcal{X}_o, \mathcal{Z}_e, \mathcal{P}^A, \mathcal{P}^B). \quad (21)$$

In this case, the posterior factorizes as

$$\begin{aligned}
P(\mathcal{X}_e, \pi | \mathcal{X}_g, \mathcal{X}_o, \mathcal{Z}_e, \mathcal{P}^A, \mathcal{P}^B) &\propto P(\mathcal{Z}_e, \mathcal{P}^B | \mathcal{X}, \pi, \mathcal{P}^A) P(\mathcal{X}_e, \pi | \mathcal{X}_g, \mathcal{X}_o, \mathcal{P}^A) \\
&\propto P(\mathcal{Z}_e | \mathcal{X}) P(\mathcal{P}^B | \mathcal{X}, \pi, \mathcal{P}^A) P(\mathcal{X}_e | \mathcal{X}_g, \mathcal{X}_o) P(\pi).
\end{aligned} \tag{22}$$

Provided that \mathcal{Z}_e are independent of the rest of the camera poses given \mathcal{X}_e , then $P(\mathcal{Z}_e | \mathcal{X}) \propto P(\mathcal{Z}_e | \mathcal{X}_e)$. Additionally, we approximate $P(\mathcal{X}_e | \mathcal{X}_g, \mathcal{X}_o)$ with the assumption that the cameras outside the temporal window do not influence the estimation of the new ones as

$$\begin{aligned}
P(\mathcal{X}_e | \mathcal{X}_g, \mathcal{X}_o) &\approx P(\mathcal{X}_e | \mathcal{X}_g) \approx \\
&P(\mathbf{x}_2^e | \mathbf{x}_1^e, \mathbf{x}_{\mathcal{X}_g}^g) P(\mathbf{x}_1^e | \mathbf{x}_{\mathcal{X}_g}^g, \mathbf{x}_{\mathcal{X}_g-1}^g) \prod_{k=0}^{\mathcal{X}_e-3} P(\mathbf{x}_{\mathcal{X}_e-k}^e | \mathbf{x}_{\mathcal{X}_e-(k+1)}^e, \mathbf{x}_{\mathcal{X}_e-(k+2)}^e)
\end{aligned} \tag{23}$$

where all the terms have been further approximated as a Markov process of second order in the same way as before.

However, operating only in a temporal window may not provide enough baseline between camera poses to capture the depth of the plane accurately. Therefore, we have enhanced the measurement set with evenly-distributed visual measurements that cover the available space explored at every iteration. In order to sparsely select sets of landmarks throughout the observed area, we project the image corners using the available camera poses and plane estimated at a given iteration. Provided that drift has not been accumulated during the estimation, we can use the projection of the image corners to determine the location of the image in the mosaic space. To this end, we compute the centroids of the reprojected corners and use K -means (Fig. 5b) to cluster the centroids into different regions of the space. Finally, we randomly pick a consecutive subset of camera motions

in each cluster to be taken into account together with their corresponding landmarks. Figure 5 depicts the proposed algorithm.

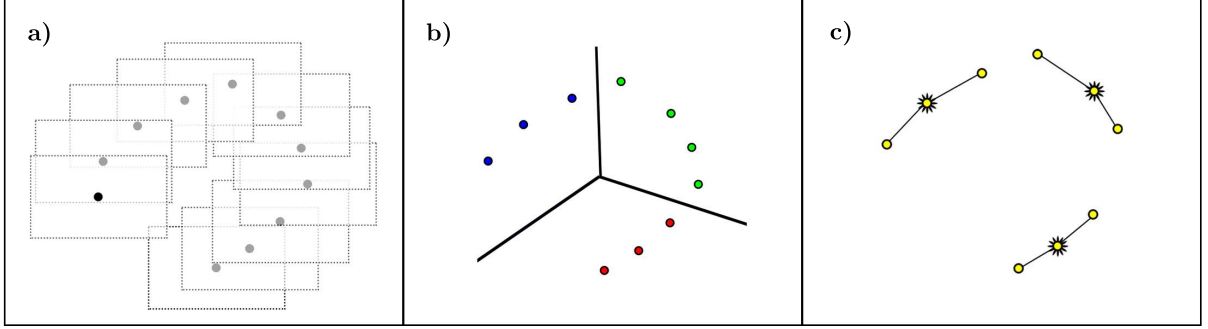


Fig. 5 a) The estimated cameras and plane produce drift-free estimates of the projected images in the mosaic space. We project the image corners (dotted contour) and compute the centroids (black dots). b) The centroids have been clustered using K -means. c) A consecutive subset of centroids has been randomly selected from each cluster. These centroids correspond to fixed camera poses as well as visual measurements, which are then used to leverage the visual pairwise relations in different areas of the space.

In a similar way as before, we can formulate the problem

$$(\hat{\mathcal{X}}_e, \hat{\pi}) = \underset{(\mathcal{X}_e, \pi)}{\operatorname{argmin}} (C_v + C_{EMT} + C_p) \quad (24)$$

where

$$C_v = \sum_{l,m \in \mathcal{W}} \sum_{i=1}^{N_{l,m}} \frac{1}{\sigma_v^2} \|\mathbf{p}^{B,i} - \mu_v^i(\mathbf{x}_l, \mathbf{x}_m, \pi, \mathbf{p}_{l,m}^{A,i})\|_2^2 \quad (25)$$

$$C_{EMT} = \sum_{k=1}^W (\mathbf{z}_{e,k} - \mathbf{x}_k^e)^T \Sigma_{EMT}^{-1} (\mathbf{z}_{e,k} - \mathbf{x}_k^e) \quad (26)$$

$$C_p = \sum_{k=3}^W (\mathbf{x}_k^e - \mu_p(\mathbf{x}_{k-1}^e, \mathbf{x}_{k-2}^e))^T \Sigma_p^{-1} (\mathbf{x}_k^e - \mu_p(\mathbf{x}_{k-1}^e, \mathbf{x}_{k-2}^e)) \quad (27)$$

$$+ (\mathbf{x}_2^e - \mu_p(\mathbf{x}_1^e, \mathbf{x}_{\chi_g}^g))^T \Sigma_p^{-1} (\mathbf{x}_2^e - \mu_p(\mathbf{x}_1^e, \mathbf{x}_{\chi_g}^g))$$

$$+ (\mathbf{x}_1^e - \mu_p(\mathbf{x}_{\chi_g}^g, \mathbf{x}_{\chi_g-1}^g))^T \Sigma_p^{-1} (\mathbf{x}_1^e - \mu_p(\mathbf{x}_{\chi_g}^g, \mathbf{x}_{\chi_g-1}^g))$$

where $\mathcal{W} \subset \mathcal{L}$ is the subset of corresponding images that has been sparsely selected as well as the ones within the temporal window. This results in much smaller non-linear least squares problems that can be solved using Gauss-Newton.

LBA produces a slightly different version of the plane in every iteration. The task of composing consistent homographies from these estimates is not trivial. For simplicity, we have opted to compose the new homographies with the estimated set of cameras and plane obtained in every iteration by using Eq. (5). This may lead to slight misregistrations in the mosaic between estimates with different planes when the camera has not explored enough area. Eventually, the estimate of the plane is going to converge and could potentially be assumed fixed, simplifying the optimization problem.

4 Experiments and Results

In this section, we introduce the algorithms to compare to, the datasets, and the metrics used to then present a suite of experiments where we evidence that our approach is drift-free, robust to EMT noise as well as robust to loss of visual information.

4.1 Algorithms, Datasets and Evaluation Metric

Algorithms: We name our proposed algorithm LBAVis+EMT. We compare it against the pairwise solution (PairVis) of the mosaicking pipeline that Brown et al.¹⁴ proposed as initialization for a further global refinement step. We also compare it against the algorithm established as the state-of-the-art in global alignment, so-called Bundle Adjustment¹⁶ (BAVis). However, rather than using homographies to parametrize the problem, we used the same parametrization as in

LBAVis+EMT to avoid over-parametrization. We also compare LBAVis+EMT against the proposed version (BAVis+EMT) of Bundle Adjustment that incorporates the EMT information.

Datasets: We introduce a synthetic (SYN, 3370 frames), a phantom-based (PHB, 902 frames), and an ex vivo human placenta (EX, 366 frames) datasets, which are composed of a set of EMT measurements as well as a sequence of images from which correspondences have been obtained using SIFT and RANSAC.¹⁴ Experimental procedures for the acquisition of the ex vivo human placenta were approved by Bloomsbury National Research Ethics Service Committee and by University College London Hospital Research and Development (REC Reference number 133888). The SYN is a xy -translation synthetic dataset in which the camera motion follows a circular pattern. It contains EMT information synthetically generated following the assumptions made on the EMT and visual information (See 2 and 3). The sequence of images was generated by selecting image regions (368x378) of a large image, representing the imaged plane, observed by the ground truth cameras. The PHB (783x782) and EX (806x779) are hand-held datasets. The PHB is recorded by imaging a printed version of a placenta taped onto a planar surface. Example images for all datasets are shown in Fig. 6.

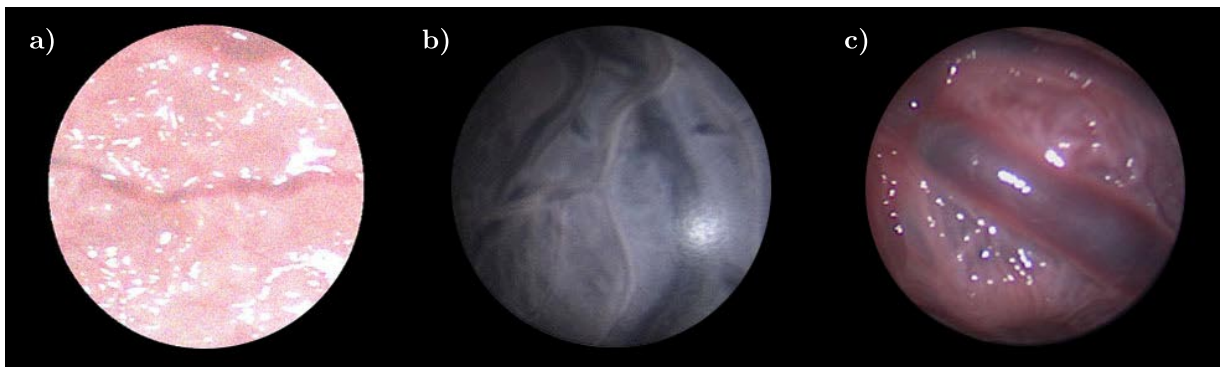


Fig. 6 a) The synthetic (SYN), b) phantom-based (PHB), and c) ex vivo human placenta (EX) datasets.

The PHB and EX datasets were recorded using the following setup: 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. According to Franz et al.,⁹ the MSE in the accuracy of the system is 0.9° in the rotation and 0.25mm in the translation in laboratory conditions. However, since the accuracy of the EMT system can vary due to external factors such as metal in the working area or position in the working volume, dynamic electromagnetic tracking errors, synchronization errors, and hand-eye calibration errors, we arbitrarily take larger⁹ standard deviations of 1° and 1mm as default values in our experiments. Synchronized video (25fps) and EMT data (40Hz) were obtained using the NifTK⁴³ software with a maximum synchronization error of 12.5ms. The fetoscope was pre-calibrated using the Matlab Camera Calibration Toolbox¹. We also pre-computed and applied the Hand-eye Calibration⁴⁴ matrix from a sequence of images of a checkerboard as well as synchronized sensor poses. The reference plane was obtained by fitting a plane to a large sweep of 3D points collected by scanning the surface with an EMT sensor. We obtained the ground truth homographies by manually registering the fetoscopic images directly to the original image of the placenta. The entire setup is shown in Fig. 7.

To provide quantitative results in terms of the accuracy of a mosaic with respect to the ground truth, we need to define the metrics.

Metrics: We parametrize a mosaic as a collection of homographies. Therefore, comparing two mosaics becomes equivalent to comparing two collections of homographies. Starting by comparing individual homographies, we define the error between any homography \mathbf{H} and the ground truth homography \mathbf{H}_{GT} as the mean residual error of a projected grid of points $\{\rho_i\}_{i=1}^{N_g} \in \Omega_I$ from the

¹<https://uk.mathworks.com/help/vision/ug/camera-calibration.html>

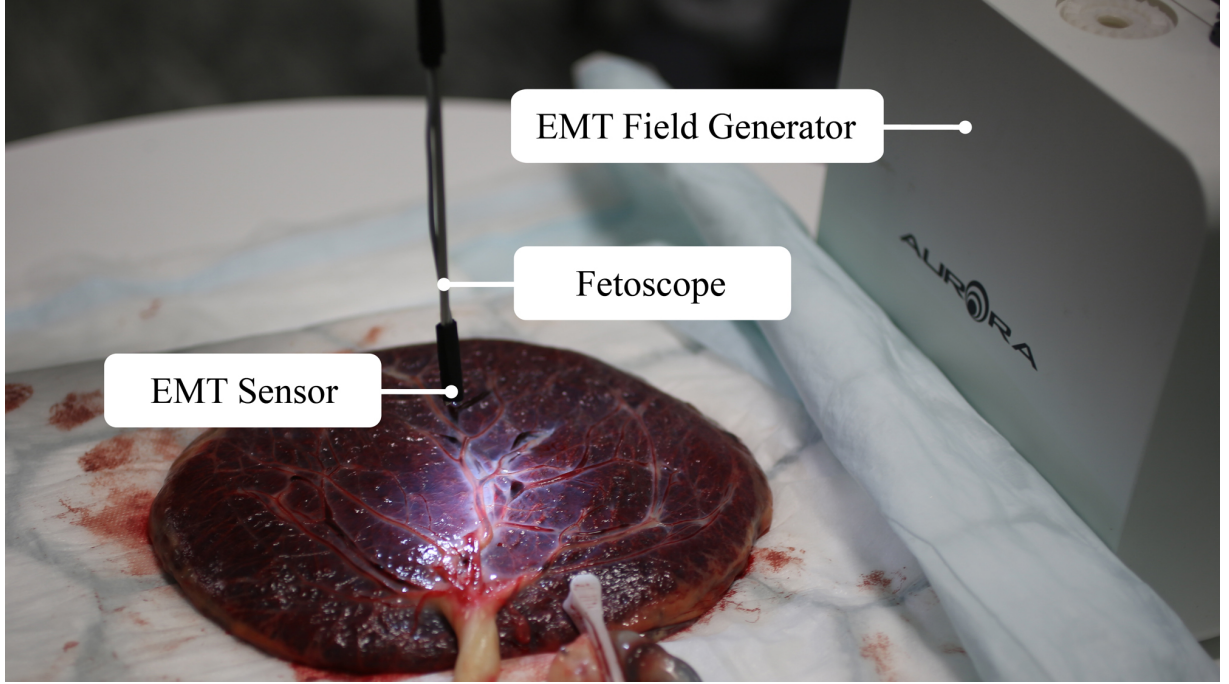


Fig. 7 The setup is composed of an EMT field generator and fetoscope in which an EMT sensor has been assembled at its tip.

image space $\Omega_I \subset \mathbb{R}^2$ to the mosaic space Ω_M

$$e_j = \frac{1}{N_g} \sum_{i=1}^{N_g} \|w(\mathbf{H}_j^{-1}, \rho_i) - w(\mathbf{H}_{GT}^{-1}, \rho_i)\|_2 \quad (28)$$

where $w(\mathbf{H}, \rho)$ projects the point ρ from the image space Ω_I to the mosaic space Ω_M through \mathbf{H} by propagating the point and converting it to Cartesian coordinates. Specifically, we have used a grid of $N_g = 100^2$ points for each comparison.

To further compare two collections of homographies, we take the mean of the error associated to each of the homographies with respect to the ground truth and define the error e_M that represents the average reprojection error in pixels

$$e_M = \frac{1}{N} \sum_{j=1}^N e_j. \quad (29)$$

4.2 Experimental Suite

In the first experiment, we show that LBAVis+EMT is not affected by long-term drift in the SYN and PHB datasets. In the second experiment, we demonstrate that the complementarity between EMT and visual information makes the system robust to the jitter effect caused by the EMT noise on the SYN dataset. In the third experiment, we evidence how the loss of visual information does not impede the creation of the mosaic in the SYN dataset. Finally, we present a set of videos for all datasets showing accurate sequential drift-free creation of the mosaics from long sequences.

The choice of parameters has been the following: $W = 5$, $\chi_e = 3$ and $K = 3$ each of which we take 5 consecutive cameras for SYN, $W = 3$, $\chi_e = 1$ and $K = 3$ each of which we take 3 consecutive cameras for PHB, and $W = 12$, $\chi_e = 6$ and $K = 3$ each of which we take 12 consecutive cameras for EX. The covariance matrix Σ_p has been estimated through an independent dataset with similar motion characteristics. Σ_{EMT} has been conservatively chosen to account for other sources of error while σ_v^2 has been set to $\sigma_v = 1$ pixel. Σ_{EMT} and Σ_p are shown in the appendix.

Once the alignment has been performed, we use a linear blending¹⁵ with a thin circular black border in each image to clearly show where it has been stitched. The use of linear blending allows for clearly distinguishing the misregistrations and it is, therefore, more interesting for demonstration purposes. However, we also use multi-band blending¹⁵ to provide more appealing results for all datasets (Video1.mp4, Video2.mp4, Video3.mp4).

4.2.1 Drift-free mosaicking

The goal of this experiment is to show that LBAVis+EMT does not drift over time. Figure 8 shows the curves for the PairVis (blue dashed), LBAVis+EMT (green), BAVis (yellow dotted)

and BAVis+EMT (green dotted) in the SYN (Fig. 8a) and PHB (Fig. 8b) datasets. The x -axis is the number of frames and the y -axis is the error e_j in pixels corresponding to a homography \mathbf{H}_j . Note that in the case of PairVis, this homography is created through a composition of pairwise homographies and therefore a point in the curve shows the cumulative error up to the corresponding frame. Both experiments show a similar trend; the growing tendency in the PairVis is expected due

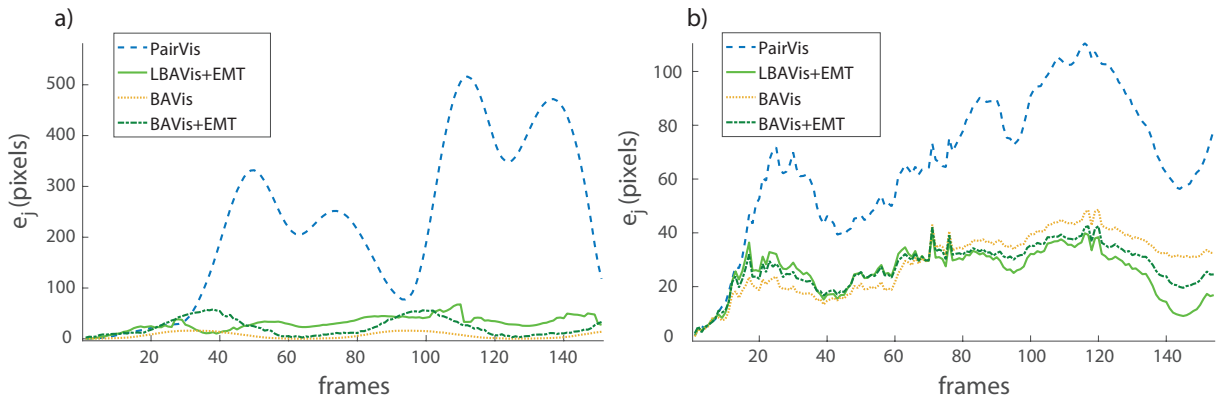


Fig. 8 Assessment of the accuracy of the PairVis (blue dashed), LBAVis+EMT (green), BAVis (yellow dotted) and BAVis+EMT (dark green dash-dotted) in a) SYN and b) PHB.

to small misregistrations in subsequent images, which leads to this long-term drift. In contrast, LBAVis+EMT maintains an approximately constant tendency over time, which demonstrates the absence of long-term drift. Additionally, the accuracy of the proposed approach is very close to that of BAVis and BAVis+EMT. This shows the feasibility of sequential methods for mosaicking in a planar scenario in terms of accuracy, when the EMT system is guiding the estimation.

We have used a relatively small number of frames (152 frames in SYN and 155 in PHB) to evidence the long-term drift in PairVis while being able to compute BAVis and BAVis+EMT. However, we emphasize that after a certain number of frames, which will depend on the quality of the pairwise registrations, the accumulated error can lead to projections of unnatural size (See Fig. 1), which may result in memory problems when creating the mosaic image. In contrast, we

Algorithm	Dataset	#Images	#Pairs	#Avg. Corr.	Matching (s)	Optimization (s)
PairVis	SYN	152	151	86	49.811	3.145
LBAVis+EMT	SYN	152	350	67	178.766	38.567
BAVis	SYN	152	1094	76	4286.259	521.819
BAVis+EMT	SYN	152	1094	76	4350.518	321.928
PairVis	PHB	155	154	222	88.192	3.333
LBAVis+EMT	PHB	155	449	184	519.447	50.638
BAVis	PHB	155	2030	105	15064.135	1488.251
BAVis+EMT	PHB	155	2030	105	15394.628	1103.748

Table 1 Runtimes for both SYN and PHB datasets for PairVis, LBAVis+EMT, BAVis and BAVis+EMT. The second column indicates the number of images in the dataset. The third column corresponds to the number of matched pairs. The fourth column is the average number of correspondences in all pairs. The fifth column is the matching runtime in seconds. Note that it includes unsuccessful matching attempts. The last column is the optimization runtime in seconds.

show in Fig. 11 how LBAVis+EMT can cope well with long sequences.

In Table 1, we show the runtimes for both SYN and PHB datasets. We report the runtimes of the steps that differ between algorithms: matching and optimization. The experiments were performed in a MacBook Pro with an Intel Core I7 at 2.5GHz with 4 cores and 16GB of RAM memory and the algorithms were implemented in Matlab using VLFeat version of SIFT and matching algorithm.⁴⁶ While PairVis is the fastest method, LBAVis+EMT shows similar performance to the gold standard Bundle Adjustment with much less computational burden. We also highlight the fact that the computational times of BAVis+EMT are smaller than BAVis. Since we provide direct, albeit noisy measurements of the latent variables to estimate, the problem is better posed. Therefore, faster convergence is expected.

4.2.2 Robustness to EMT noise

Inherently, the use of EMT information produces a jitter effect in the mosaic due to the noise in the camera pose measurements. The goal of this experiment is to assess the accuracy of LBAVis+EMT and show that when EMT information is fused with visual information, the jitter effect is mitigated in the creation of the mosaic. For this purpose, we created seven synthetic datasets, each with

different EMT noise statistics. We have denoted the scalar ν to be the standard deviation of the EMT noise in the camera poses. We chose a small subset of only 17 frames since we are now only interested in the quality of pairwise registrations.

For every SYN dataset, we assessed the accuracy of the resulting mosaic for LBAVis+EMT and compared it against the jittery baseline composition of the mosaic using only EMT information and the ground truth plane. Figure 9a displays the graphs for both algorithms. The x -axis represents the different datasets from best to worse EMT noise statistics ν while the y -axis corresponds to the average error in the mosaic e_M measured in pixels. We can see how while the EMT-based compo-

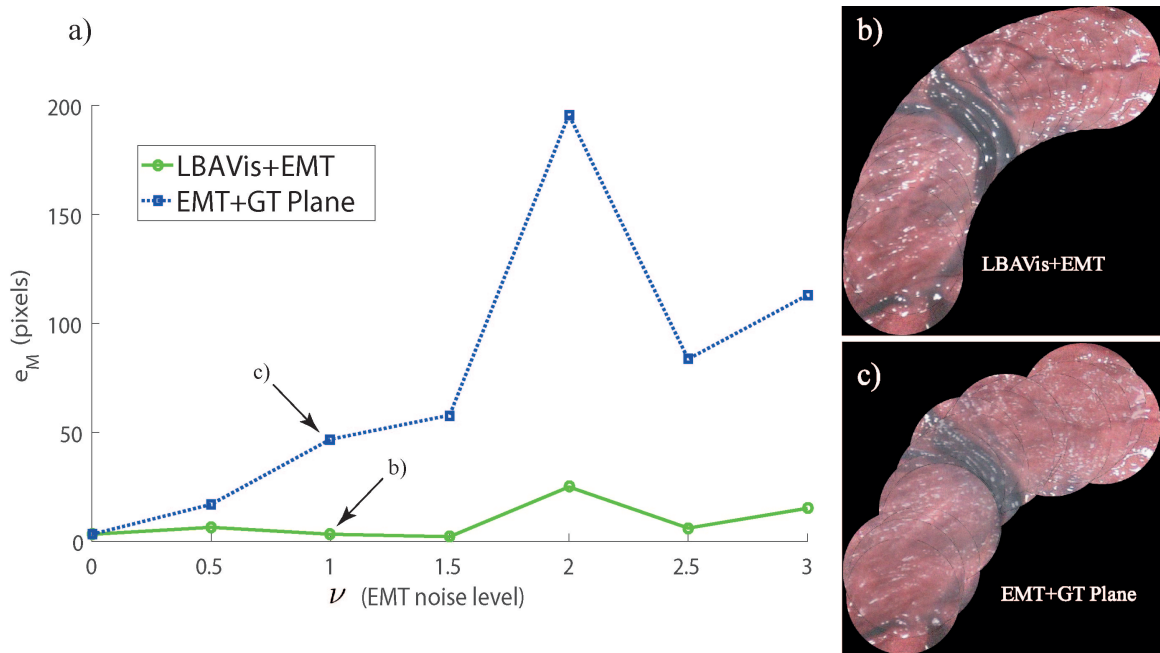


Fig. 9 a) In the x -axis, the multiplier ν defines the standard deviation of the rotation and translation in degrees and millimeters respectively of seven versions of the SYN dataset. In the y -axis, the error in pixels e_M in the mosaic. Every point in the graph represents the error in a mosaic created by using either LBAVis+EMT (green) or EMT (blue dotted) with the ground truth plane (EMT+GT Plane). b) LBAVis+EMT mosaic for $\nu = 1$. c) EMT+GT Plane mosaic for $\nu = 1$.

sition shows an approximately linear tendency with the increase of ν , LBAVis+EMT outperforms

it by showing an approximately constant accuracy. This stands to reason since, after $\nu = 1$, the quality of the EMT information is really bad and thus barely used. Then, the visual information plays a major role in the estimation. We have highlighted the case $\nu = 1$ and displayed the mosaics corresponding to $\nu = 1$ since it is the value of choice in other experiments. Figure 9b shows an accurate mosaic created with LBAVis+EMT while Fig. 9c shows the jitter effect of EMT information in the mosaic.

4.2.3 Robustness to sudden loss of visual information

In this experiment, we test our approach when no corresponding landmarks can be found between pairs of frames. We created an alternative dataset in the same manner than the SYN dataset was created. This new dataset contains 62 frames of which 12 randomly selected ones were replaced by black frames to simulate lack of visual content. The black frames are: 7, 11, 12, 23, 24, 37, 38, 42, 43, 45, 51 and 54. Then, we ran LBAVis+EMT to assess its accuracy with missing visual information. Figure 10a shows how despite loss of visual information, when PairVis would fail, LBAVis+EMT has been able to successfully create a mosaic. Black circles in the mosaic (Fig. 10a) represent the reprojection of the missing frames in the mosaic space. Figures 10b and 10c show the estimated plane and six components of the camera pose (green); rotation (r_x, r_y, r_z) and translation (t_x, t_y, t_z) , the EMT measurements (blue crosses) and their ground truth (red) respectively for each frame. We convert the plane to azimuth, elevation, and distance for an easier interpretation and detail the conversion in the appendix. Therefore, when no landmarks are available, the estimation can continue successfully and provide a reasonable estimate of the camera poses and plane.

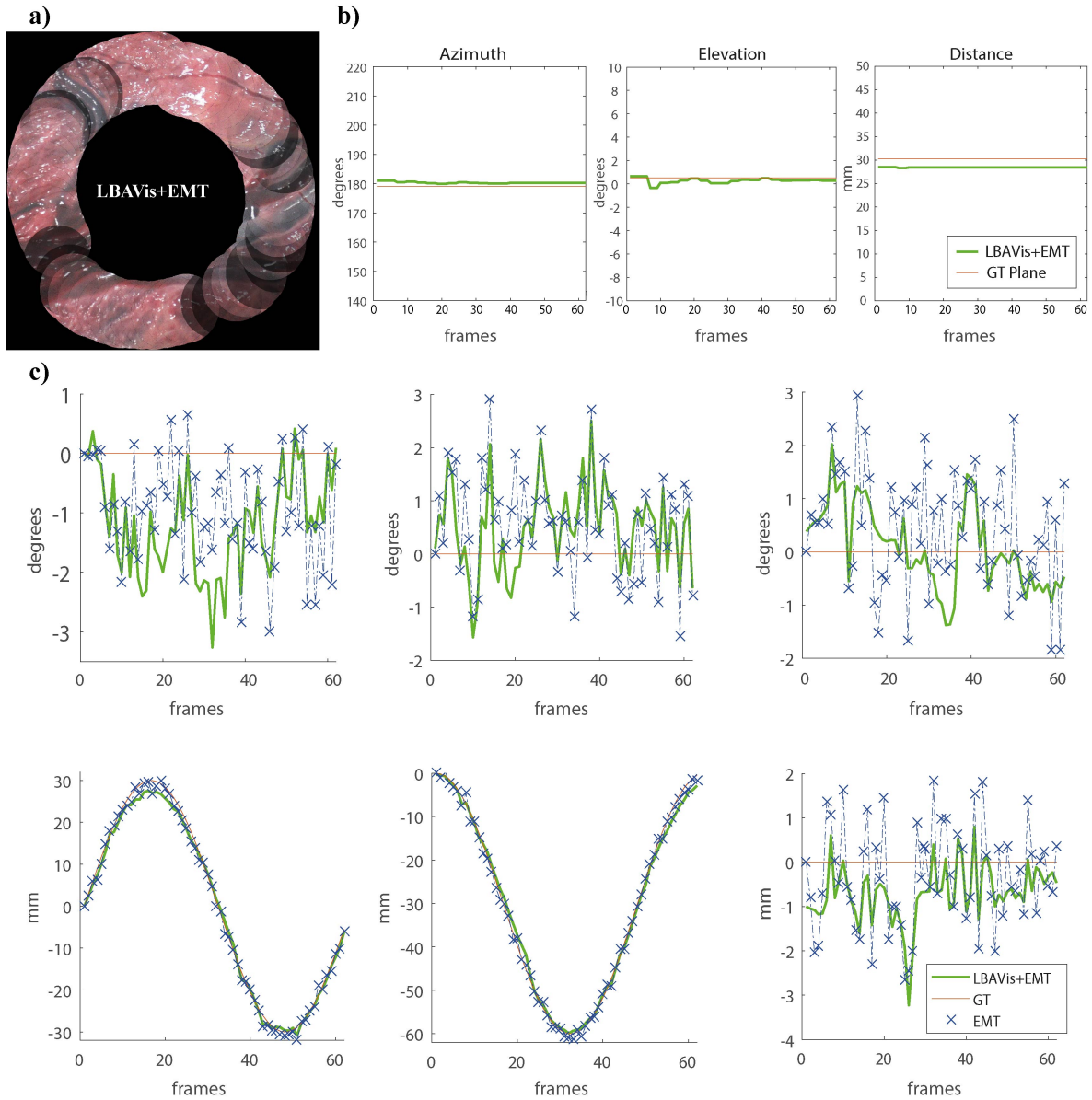


Fig. 10 a) LBAVis+EMT mosaic of a synthetic dataset where some random frames have been replaced by black frames to simulate the lack of visual content (7, 11, 12, 23, 24, 37, 38, 42, 43, 45, 51 and 54). Black circles in the mosaic represent the reprojection of the missing frames in the mosaic space. b) Estimation of the plane (green) and ground truth (red). c) Estimation of the camera pose (green); rotation (r_x, r_y, r_z) and translation (t_x, t_y, t_z), EMT measurements (blue crosses) and ground truth (red).

4.2.4 Sequential creation and blending of the mosaics

Figures 11a, b, and c show the results of running LBAVis+EMT in the SYN (3770 frames), PHB (902 frames), and EX (366 frames) datasets respectively. Figure 12 shows the graph corresponding

to the error e_j in every frame of the SYN dataset. We provide videos (Video 1, MP4, 5.1MB, Video 2, MP4, 3.4MB, Video 3, MP4, 3.1MB) that illustrate the sequential creation of the mosaics by showing the subsequent blending of every new image into the mosaic image. We recommend the reader visualize the videos for a better understanding of the results. In this experiment, the goal is to demonstrate that our approach can create accurate long mosaics in a sequential fashion.

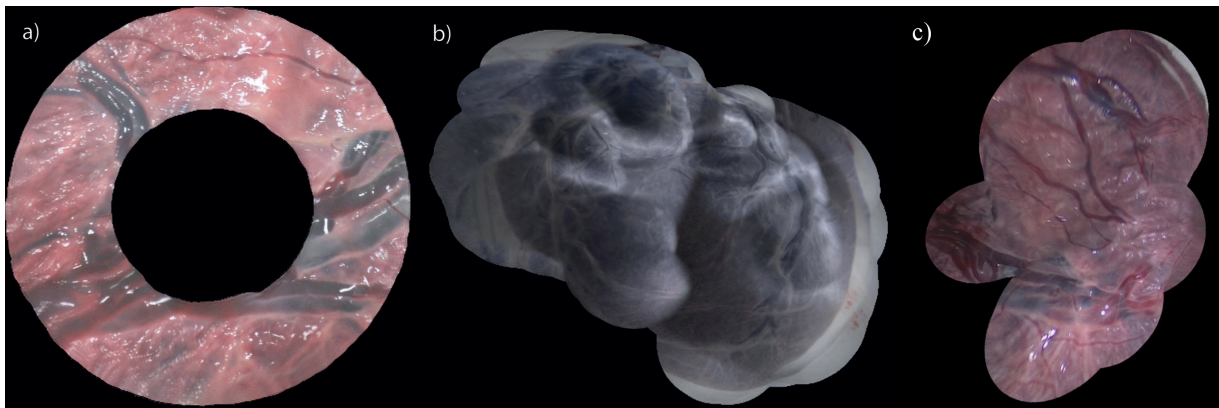


Fig. 11 Mosaic using LBAVis+EMT in the a) SYN (Video 1, MP4, 5.1MB), b) PHB (Video 2, MP4, 3.4MB), and c) EX (Video 3, MP4, 3.1MB) datasets.

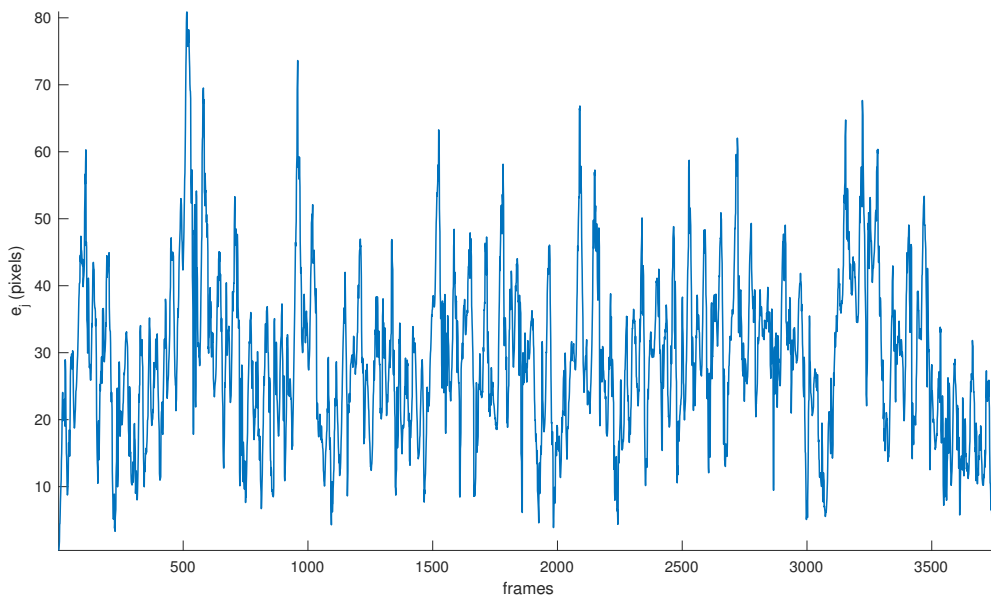


Fig. 12 Error e_j in pixels of LBAVis+EMT in the SYN dataset.

5 Discussion

Our results confirm that the fusion between the EMT and visual information using the proposed probabilistic model in a sequential fashion does not accumulate drift. However, there exists an error within the acceptable range of camera poses in which estimates can lay. We believe that the major cause of this error is the use of previous estimates as fixed camera poses, which encourages continuity on subsequent estimations. The effect of this error can be clearly appreciated in Fig. 11a. In contrast to the PairVis, which in Fig. 8 has drifted after approximately 20 frames, our approach is able to create a consistent mosaic after 3770 frames. However, a fixed point in one of the circular loops does not necessarily match to the exact same point when the scene is revisited (Fig. 11a). The range of error experienced corresponds to the spread where the EMT system allows the estimates to be as long as there is pairwise consistency. To further highlight this fact, we can see that when the estimation exceeds the range of error allowed by the model, an occasional pull towards the true value can be observed (e.g. see frames 18-19). We did not include a version of the Local Bundle Adjustment using only visual information in our results since it would still use exclusively pairwise visual measurements and therefore, it would drift.

When the scene is revisited, images are not necessarily stitched in an exactly consistent location. This is a limitation of our approach, we do not constrain the revisited positions to match. However, an immediate extension that can tackle this problem would consist of the use of a spatiotemporal window to consider also regions of the space that are being revisited. While the probabilistic formulation would remain valid, the spatial window would optimize for loop closures, yet avoiding the computational cost associated with re-estimating the cameras within the loop, since we would rely on the EMT information to situate the loop roughly in a correct location from the

beginning.

Our results show how, despite the inherent noise in the EMT system, its measurements can be used to produce accurate pairwise registrations. However, since independence between corresponding points has been assumed in the model and the number of points can be high, if not dealt with, it translates into underestimated uncertainty in the visual information, thus leading to less reliance on the EMT information. Nonetheless, our setup does not yet fully simulate clinical images; the matching process in our datasets is in general easier than in clinical videos, with an increased number of matches, which accentuates the imbalance between both modalities.

The placenta is not completely planar, the violation of this assumption will inevitably produce misregistration errors in those areas where the non-planarities are more prominent. Figure 11c demonstrates that these errors are small enough to consider the assumption of planarity valid in an ex vivo scenario. However, further research should be conducted on in vivo data.

Additionally, we see how the feature-based and outlier removal strategy do not perform optimally in ex vivo tissue (Fig. 11c), which leads to misregistrations in the mosaic. We believe that further research in registration of pairwise images must be done for ex vivo and in vivo data. However, this is out of the scope of this paper.

If we closely analyze how the noise in the corresponding points is propagated, we see that there are two error terms playing a role in a point $\mathbf{p}^{i,B}$ in image B; the corresponding point $\mathbf{p}^{i,A}$ in image A gets propagated through the homography, which makes one error term highly correlated between points, and another one is additive, i.e. $\mathbf{p}^{i,B} = \mathbf{H}(\mathbf{p}^{i,A} + \epsilon^{i,A}) + \epsilon^{i,B}$. Therefore, the distribution already undergoes a non-linear function, and there can be other non-linear factors such as distortion errors that can complicate even more the resulting distribution. For this reason, we opted to approximate such distribution as a Gaussian, which works well in practice.

We have also demonstrated how our approach can perform adequately even when visual information is missing. The reason for this is that we have strong information about missing camera poses: For each empty frame, we have an EMT measurement that tells us the approximate location of the camera, and a temporal model that also provides prior knowledge of its location. Furthermore, the fact that some images are missing does not impede the estimation of the plane as long as the visual information has enough baseline between images.

Our approach estimates the distance d from the reference camera to the surface, which is directly related to the distance between the camera and the surface. This distance can be a powerful cue for the surgeon since photocoagulation must be done at a certain fixed distance from the placenta. However, neither the EMT nor visual information in isolation can estimate such distance. Actually, if no EMT information is provided, the distance becomes a free parameter due to the inherent scale ambiguity in monocular cameras. For this reason, its estimation is not straightforward and further assessment of the accuracy needs to be conducted. Additionally, the Hand-eye calibration is in general problematic; a set of images of a known object, e.g. a checkerboard, must be taken with the fetoscope and the attached EMT sensor in controlled conditions. One may also need to apply some heuristics to achieve an acceptable accuracy.

6 Conclusions

We have presented a probabilistic model for robust drift-free sequential mosaicking that fuses imagery and data from an EMT system in the case where a planar or quasi-planar object is imaged in a hand-held motion. We have shown that our method does not accumulate error; a problem that affects all monocular pairwise mosaicking systems which use exclusively visual information. Therefore, we have been able to create long mosaics obtaining an accuracy comparable to

the state-of-the-art Bundle Adjustment. In spite of the inherent noise in EMT systems, we have demonstrated that our approach can still generate accurate mosaics while leveraging its guidance. Furthermore, we have shown its feasibility even when there is a loss of visual information.

In terms of future work, we are considering the following research lines: (i) Clinical fetoscopic datasets present a challenge in terms of obtaining a set of valid matches. In this work, we have not addressed this problem; however, this is clearly a limitation towards fetoscopic mosaicking that we plan to address. (ii) In order to eliminate the error within the EMT bounds, we could consider the use of Weighted Local Bundle Adjustment (WLBA). This would soften the effect of fixed cameras and reconsider already estimated cameras according to their uncertainty in the current estimation. (iii) To further avoid a normalization in the corresponding points, modeling the fact that the errors in the points are correlated is an interesting idea, which could avoid the normalization process. (iv) The inclusion of the visual and Hand-eye matrix in the model is an attractive research line, which if successful, would remove a tedious step in the operating room. (v) Last, this study serves as a proof of concept to a future real-time version of the approach. We believe this is possible given that algorithms with similar computational load have been successfully implemented in real-time.

Given the low quality of fetoscopic images, we believe that the inclusion of the EMT system in the mosaicking process is fundamental to achieve a robust and accurate mosaic, independently of the number of frames.

Disclosures

We have nothing to disclose.

Acknowledgments

This work was supported by an Innovative Engineering for Health award by Wellcome Trust [WT101957]; Engineering and Physical Sciences Research Council (EPSRC) [NS/A000027/1]. Jan Deprest is being funded by the Great Ormond Street Hospital Charity. Sebastien Ourselin receives funding from EPSRC (EP/H046410/1, EP/J020990/1, EP/K005278) and the MRC (MR/J01107X/1). Danail Stoyanov receives funding from the EPSRC (EP/N013220/1, EP/N022750/1, EP/N027078/1, NS/A000027/1), The Wellcome Trust (WT101957, 201080/Z/16/Z) and the EU-Horizon2020 project EndoVESPA (H2020-ICT-2015-688592). Juan Eugenio Iglesias receives funding from ERC Starting Grant agreement No 677697 (project BUNGEE-TOOLS). Marcel Tella-Amo is supported by the EPSRC-funded UCL Centre for Doctoral Training in Medical Imaging (EP/L016478/1). We would also like to thank Jyotirmoy Banerjee for the useful discussions as well as Efthymios Maneas for the contribution of the ex vivo placenta.

References

- 1 L. Lewi, L. Gucciardo, T. Van Mieghem, *et al.*, “Monochorionic diamniotic twin pregnancies: natural history and risk stratification,” *Fetal diagnosis and therapy* **27**(3), 121–133 (2010). [doi:10.1159/000313300].
- 2 A. Baschat, R. H. Chmait, J. Deprest, *et al.*, “Twin-to-twin transfusion syndrome (ttts),” *Journal of perinatal medicine* **39**(2), 107–112 (2011).
- 3 F. Slaghekke, E. Lopriore, L. Lewi, *et al.*, “Fetoscopic laser coagulation of the vascular equator versus selective coagulation for twin-to-twin transfusion syndrome: an open-label ran-

- domised controlled trial,” *The Lancet* **383**(9935), 2144–2151 (2014). [doi:10.1016/S0140-6736(13)62419-8].
- 4 M.-V. Senat, J. Deprest, M. Boulvain, *et al.*, “Endoscopic laser surgery versus serial amnioreduction for severe twin-to-twin transfusion syndrome,” *New England Journal of Medicine* **351**(2), 136–144 (2004).
 - 5 P. Daga, F. Chadebecq, D. Shakir, *et al.*, “Real-time mosaicing of fetoscopic videos using sift,” in *Proceedings of SPIE*, **9786**, SPIE Medical Imaging (2016).
 - 6 M. Reeff, *Mosaicing of endoscopic placenta images*, Hartung-Gorre (2011).
 - 7 M. Tella-Amo, P. Daga, F. Chadebecq, *et al.*, “A combined em and visual tracking probabilistic model for robust mosaicking: Application to fetoscopy,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 84–92 (2016).
 - 8 P. T. Tran, P.-L. Chang, H. De Praetere, *et al.*, “3d catheter shape reconstruction using electromagnetic and image sensors,” *Journal of Medical Robotics Research* , 1740009 (2017).
 - 9 A. M. Franz, T. Haidegger, W. Birkfellner, *et al.*, “Electromagnetic tracking in medicine: a review of technology, validation, and applications,” *IEEE transactions on medical imaging* **33**(8), 1702–1725 (2014). [doi:10.1109/TMI.2014.2321777].
 - 10 K. Cleary, H. Zhang, N. Glossop, *et al.*, “Electromagnetic tracking for image-guided abdominal procedures: Overall system and technical issues,” in *2005 IEEE Engineering in Medicine and Biology 27th Annual Conference*, 6748–6753, IEEE (2005).
 - 11 B. J. Wood, H. Zhang, A. Durrani, *et al.*, “Navigation with electromagnetic tracking for interventional radiology procedures: a feasibility study,” *Journal of vascular and interventional radiology* **16**(4), 493–505 (2005). [doi:10.1097/01.RVI.0000148827.62296.B4].

- 12 K. Vyas, M. Hughes, and G.-Z. Yang, “Electromagnetic tracking of handheld high-resolution endomicroscopy probes to assist with real-time video mosaicking,” in *SPIE BiOS*, 93040Y–93040Y, International Society for Optics and Photonics (2015).
- 13 A. Dore, G. Smoljkic, E. Vander Poorten, *et al.*, “Catheter navigation based on probabilistic fusion of electromagnetic tracking and physically-based simulation,” in *Intelligent Robots and Systems (IROS), 2012 IEEE/RSJ International Conference on*, 3806–3811, IEEE (2012).
- 14 M. Brown, D. G. Lowe, *et al.*, “Recognising panoramas.,” in *ICCV*, **3**, 1218 (2003).
- 15 R. Szeliski, “Image alignment and stitching: A tutorial,” *Foundations and Trends® in Computer Graphics and Vision* **2**(1), 1–104 (2006).
- 16 H.-Y. Shum and R. Szeliski, “Construction of panoramic image mosaics with global and local alignment,” in *Panoramic vision*, 227–268, Springer (2001).
- 17 E. Michaelsen, “Stitching large maps from videos taken by a camera moving close over a plane using homography decomposition,” in *ISPRS Conference (CD), PIA 2011*, **38**(3) (2011).
- 18 H. S. Sawhney, S. Hsu, and R. Kumar, “Robust video mosaicing through topology inference and local to global alignment,” in *European conference on computer vision*, 103–119, Springer (1998).
- 19 R. Mur-Artal, J. M. M. Montiel, and J. D. Tardos, “ORB-SLAM: A Versatile and Accurate Monocular SLAM System,” *IEEE Trans. Robot.* **31**(5), 1147–1163 (2015). [doi:10.1109/TRO.2015.2463671].
- 20 F. Caballero, L. Merino, J. Ferruz, *et al.*, “Improving vision-based planar motion estima-

- tion for unmanned aerial vehicles through online mosaicing,” *Proc. - IEEE Int. Conf. Robot. Autom.* **2006**(May), 2860–2865 (2006).
- 21 H. Strasdat, J. M. M. Montiel, and A. J. Davison, “Scale Drift-Aware Large Scale Monocular SLAM,” *Robot. Sci. Syst.* **2**, 5 (2010). [doi:10.15607/RSS.2010.VI.010].
- 22 J. Civera, A. J. Davison, J. A. Magallón, *et al.*, “Drift-free real-time sequential mosaicing,” *International Journal of Computer Vision* **81**(2), 128–137 (2009). [doi:10.1007/s11263-008-0129-5].
- 23 S. Lovegrove, *Parametric dense visual SLAM*, Imperial College London (2012).
- 24 E. Malis and S. Benhimane, “Homography-based 2D Visual Tracking and Servoing,” *Int. J. Rob. Res.* **26**(7), 661–676 (2007). [doi:10.1177/0278364907080252].
- 25 F. Caballero, L. Merino, J. Ferruz, *et al.*, “A visual odometer without 3D reconstruction for aerial vehicles. Applications to building inspection,” *Proc. - IEEE Int. Conf. Robot. Autom.* **2005**(April), 4673–4678 (2005).
- 26 C. Olsson and A. Eriksson, “Triangulating a plane,” in *Scandinavian Conference on Image Analysis*, 13–23, Springer (2011). [doi:10.1007/978-3-642-21227-7-2].
- 27 S. Atasoy, D. P. Noonan, S. Benhimane, *et al.*, “A global approach for automatic fibroscopic video mosaicing in minimally invasive diagnosis,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 850–857, Springer (2008). [doi:10.1007/978-3-540-85988-8-101].
- 28 A. Eudes, S. Naudet-Collette, M. Lhuillier, *et al.*, “Weighted local bundle adjustment and application to odometry and visual slam fusion,” (2010).

- 29 M. Agrawal and K. Konolige, “Real-time localization in outdoor environments using stereo vision and inexpensive gps,” in *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, **3**, 1063–1068, IEEE (2006). [doi:10.1109/ICPR.2006.962].
- 30 T. Hamel, R. Mahony, J. Trumpf, *et al.*, “Homography estimation on the special linear group based on direct point correspondence,” in *Decision and Control and European Control Conference (CDC-ECC), 2011 50th IEEE Conference on*, 7902–7908, IEEE (2011). [doi:10.1109/CDC.2011.6160724].
- 31 S. Leutenegger, S. Lynen, M. Bosse, *et al.*, “Keyframe-based visual-inertial odometry using nonlinear optimization,” *The International Journal of Robotics Research* **34**(3), 314–334 (2015). [doi:10.1177/0278364914554813].
- 32 G. Klein and T. Drummond, “Tightly integrated sensor fusion for robust visual tracking,” *Image Vis. Comput.* **22**(10), 769–776 (2004). [doi:10.1016/j.imavis.2004.02.007].
- 33 D. Strelow and S. Singh, “Motion estimation from image and inertial measurements,” *The International Journal of Robotics Research* **23**(12), 1157–1195 (2004).
- 34 Z. Xian, J. Lian, M. Shan, *et al.*, “A square root unscented kalman filter for multiple view geometry based stereo cameras/inertial navigation,” *International Journal of Advanced Robotic Systems* **13**(5), 1729881416664850 (2016). [doi:10.1177/1729881416664850].
- 35 R. Hartley and A. Zisserman, “Multiple view geometry in computer vision,” *Robotica* **23**(2), 271–271 (2005).
- 36 D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” *International journal of computer vision* **60**(2), 91–110 (2004).

- 37 H. Bay, A. Ess, T. Tuytelaars, *et al.*, “Speeded-up robust features (surf),” *Computer vision and image understanding* **110**(3), 346–359 (2008).
- 38 E. Rublee, V. Rabaud, K. Konolige, *et al.*, “Orb: An efficient alternative to sift or surf,” in *Computer Vision (ICCV), 2011 IEEE international conference on*, 2564–2571, IEEE (2011).
- 39 S. J. Prince, *Computer vision: models, learning, and inference*, Cambridge University Press (2012). [doi:10.1017/CBO9780511996504].
- 40 T. Haidegger, G. Fenyvesi, B. Sirokai, *et al.*, “Towards unified electromagnetic tracking system assessment-static errors,” in *Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE*, 1905–1908, IEEE (2011).
- 41 E. Wilson, Z. Yaniv, H. Zhang, *et al.*, “A hardware and software protocol for the evaluation of electromagnetic tracker accuracy in the clinical environment: a multi-center study,” in *Medical Imaging 2007: Visualization and Image-Guided Procedures*, **6509**, 65092T, International Society for Optics and Photonics (2007).
- 42 A. J. Davison, I. D. Reid, N. D. Molton, *et al.*, “Monoslam: Real-time single camera slam,” *IEEE transactions on pattern analysis and machine intelligence* **29**(6), 1052–1067 (2007).
- 43 M. J. Clarkson, G. Zombori, S. Thompson, *et al.*, “The niftk software platform for image-guided interventions: platform overview and niftylink messaging,” *International journal of computer assisted radiology and surgery* **10**(3), 301–316 (2015). [doi:10.1007/s11548-014-1124-7].
- 44 K. Pachtrachai, M. Allan, V. Pawar, *et al.*, “Hand-eye calibration for robotic assisted minimally invasive surgery without a calibration object,” in *Intelligent Robots and Systems (IROS), 2016 IEEE/RSJ International Conference on*, 2485–2491, IEEE (2016).

45 N. D. Glossop, “Advantages of optical compared with electromagnetic tracking,” *JBJS* **91**(Supplement_1), 23–28 (2009). [doi:10.2106/JBJS.H.01362].

46 A. Vedaldi and B. Fulkerson, “VLFeat: An open and portable library of computer vision algorithms.” <http://www.vlfeat.org/> (2008).

Marcel Tella-Amo is a Ph.D. student at University College of London (UCL). He also received his MRes in Medical Imaging at UCL (2016). Previously, he received his BE+MSc in Telecom Engineering focused on image processing at Technical University of Catalonia (UPC), Barcelona. His research interests include machine learning and computer vision, specifically mosaicking techniques and SLAM.

Loic Peter is a postdoctoral research associate at the University College London (UCL). He received his Ph.D. in computer science from the Technical University of Munich (TUM), Germany. Previous to that, he received a French engineering degree from the Ecole Centrale Paris in 2011 and his MSc from the Ecole 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 at the Chair for Computer Aided Medical Procedures & Augmented Reality (CAMPAR) of the Technical University of Munich (TUM). He received his MSc in Computational Science and Engineering from TUM as well. Dzhoshkun’s main research interest is intra-operative imaging. His work focuses on software and hardware engineering solutions for real-time medical imaging technologies.

Jan Deprest is an obstetrician-gynaecologist, 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 University College of London (UCL). He received his Ph.D. in computer science at Imperial College London specializing in medical image computing. Previous to that, he studied electronics and computer systems engineering at King's College before completing. My research interests and expertise are in surgical vision and computational imaging, surgical robotics, image guided therapies and surgical process analysis

Juan Eugenio Iglesias is a senior research fellow at University College of London (UCL). Previously, he performed post-doctoral research at BCBL (San Sebastian) and Martinos Center for Biomedical Imaging, MGH and Harvard Medical School. He received his Ph.D. in Biomedical Imaging at UCLA. He obtained his MSc in Electrical Engineering at KTH (Stockholm) and MSc in Telecom Engineering at University of Sevilla. His research interests include brain MRI segmentation and generative models of MRI data.

Tom Vercauteren is a Senior Lecturer at University College London (UCL) and Deputy Director of the Wellcome / EPSRC Centre for Surgical and Interventional Sciences. He received his Ph.D. at the Asclepios Project-Team in Inria Sophia Antipolis from Ecole des Mines de Paris. He obtained his MSc in Electrical Engineering at Columbia University and graduated from Ecole Polytechnique. He has also been working for Mauna Kea Technologies.

Sebastien Ourselin is Director of the Wellcome / EPSRC Centre for Surgical and Interventional Sciences. He is Vice-Dean (Health) at the Faculty of Engineering Sciences (UCL), Director of the Institute of Healthcare Engineering and of the EPSRC Centre for Doctoral Training in Medical Imaging, Head of the Translational Imaging Group within CMIC, and Head of Image Analysis at

the DRC. His research interests include image registration, segmentation, statistical shape modeling, surgical simulation, image-guided therapy and minimally invasive surgery.

7 Appendix

7.1 Covariance matrices

$$\Sigma_{EMT} = \text{diag}([0.017, 0.017, 0.017, 1, 1, 1]) \quad (30)$$

$$\Sigma_p = \text{diag}([0.0044, 0.0044, 0.0044, 15.2178, 15.2178, 15.2178])$$

where $\text{diag}(\cdot)$ denotes a diagonal matrix.

7.2 Conversion from Cartesian coordinates to azimuth, elevation and distance

$$\begin{aligned} \tan(az) &= \frac{y}{z} \\ \tan(elev) &= \frac{x}{\sqrt{z^2 + y^2}} \\ d &= \sqrt{z^2 + y^2 + x^2} \end{aligned} \quad (31)$$

List of Figures

- 1 Mosaics of 200 synthetic images where the camera has moved following a circular pattern of four laps. a) The ground truth mosaic. b) A mosaic with drift due to the accumulation of error in subsequent iterations.

- 2 A virtual camera is placed at the origin of an arbitrary global coordinate system. The relation between images and cameras is defined in Eqs. (3) and (5).
- 3 Schematic of the nomenclature of the correspondences. $\mathcal{P}_{k+1,k}^A$ (green empty circles) is the set of points of image A for which $N_{k+1,k} = 4$ correspondences from image k to $k + 1$ are available, being $\mathcal{P}_{k+1,k}^B$ (green colored circles) the corresponding landmarks in image B. $\mathcal{P}_{k+2,k}^A$ (blue empty triangles) is the set of points of image A for which $N_{k+2,k} = 5$ correspondences from image k to $k + 2$ are available, being $\mathcal{P}_{k+2,k}^B$ (blue colored triangles) the corresponding landmarks in image B. As an example, the landmark $\mathbf{p}_{k+1,k}^{A,1}$ corresponds to the point $\mathbf{p}_{k+1,k}^{B,1}$ and the point $\mathbf{p}_{k+2,k}^{A,5}$ corresponds to the point $\mathbf{p}_{k+2,k}^{B,5}$.
- 4 Graphical model of the proposed probabilistic framework. Circles represent random variables, which can be either latent (white background) or observed (shaded). Parameters of the model are depicted as a point.
- 5 a) The estimated cameras and plane produce drift-free estimates of the projected images in the mosaic space. We project the image corners (dotted contour) and compute the centroids (black dots). b) The centroids have been clustered using K -means. c) A consecutive subset of centroids has been randomly selected from each cluster. These centroids correspond to fixed camera poses as well as visual measurements, which are then used to leverage the visual pairwise relations in different areas of the space.
- 6 a) The synthetic (SYN), b) phantom-based (PHB), and c) ex vivo human placenta (EX) datasets.

- 7 The setup is composed of an EMT field generator and fetoscope in which an EMT sensor has been assembled at its tip.
- 8 Assessment of the accuracy of the PairVis (blue dashed), LBAVis+EMT (green), BAVis (yellow dotted) and BAVis+EMT (dark green dash-dotted) in a) SYN and b) PHB.
- 9 a) In the x -axis, the multiplier ν defines the standard deviation of the rotation and translation in degrees and millimeters respectively of seven versions of the SYN dataset. In the y -axis, the error in pixels e_M in the mosaic. Every point in the graph represents the error in a mosaic created by using either LBAVis+EMT (green) or EMT (blue dotted) with the ground truth plane (EMT+GT Plane). b) LBAVis+EMT mosaic for $\nu = 1$. c) EMT+GT Plane mosaic for $\nu = 1$.
- 10 a) LBAVis+EMT mosaic of a synthetic dataset where some random frames have been replaced by black frames to simulate the lack of visual content (7, 11, 12, 23, 24, 37, 38, 42, 43, 45, 51 and 54). Black circles in the mosaic represent the reprojection of the missing frames in the mosaic space. b) Estimation of the plane (green) and ground truth (red). c) Estimation of the camera pose (green); rotation (r_x, r_y, r_z) and translation (t_x, t_y, t_z) , EMT measurements (blue crosses) and ground truth (red).
- 11 Mosaic using LBAVis+EMT in the a) SYN (Video 1, MP4, 5.1MB), b) PHB (Video 2, MP4, 3.4MB), and c) EX (Video 3, MP4, 3.1MB) datasets.
- 12 Error e_j in pixels of LBAVis+EMT in the SYN dataset.