A Combined EM and Visual Tracking Probabilistic Model for Robust Mosaicking: Application to Fetoscopy.

Marcel Tella[1], Pankaj Daga[1], François Chadebecq[1,2], Stephen Thompson[1], Dzhoshkun I. Shakir[1], George Dwyer[1,2], Ruwan Wimalasundera[4], Jan Deprest[1,3], Danail Stoyanov[2], Tom Vercauteren[1], and Sebastien Ourselin[1]

[1]Translational Imaging Group, CMIC, University College London, UK
[2]Surgical Robot Vision Group, CMIC, University College London, UK
[3]University Hospitals Leuven, Department of Obstetrics and Gynaecology, Leuven, Belgium
[4]University College Hospital, UK

Abstract

Twin-to-Twin Transfusion Syndrome (TTTS) is a progressive pregnancy complication in which inter-twin vascular connections in the shared placenta result in a blood flow imbalance between the twins. The most effective therapy is to sever these connections by laser photo-coagulation. However, the limited field of view of the fetoscope hinders their identification. A potential solution is to augment the surgeon's view by creating a mosaic image of the placenta. State-of-the-art mosaicking methods use feature-based approaches, which have three main limitations: (i) they are not robust against corrupt data e.g. blurred frames, (ii) temporal information is not used, (iii) the resulting mosaic suffers from drift. We introduce a probabilistic temporal model that incorporates electromagnetic and visual tracking data to achieve a robust mosaic with reduced drift. By assuming planarity of the imaged object, the nRT decomposition can be used to parametrize the state vector. Finally, we tackle the non-linear nature of the problem in a numerically stable manner by using the Square Root Unscented Kalman Filter. We show an improvement in performance in terms of robustness as well as a reduction of the drift in comparison to state-of-the-art methods in synthetic, phantom and ex vivo datasets.

1. Introduction

Twin-to-Twin Transfusion Syndrome is a progressive complication of monochorionic diamniotic (MCDA) pregnancies. Inter-twin vascular connections shared in the placenta result in an imbalance in the blood circulation which can lead to the death of both twins [5, 12]. Furthermore, cardiac complications may arise in one of the fetuses due to the excess of blood whereas the other may suffer from anemia, an abnormal decrease of the hemoglobin in the blood.

The recommended treatment for TTTS is laser photocoagulation. This involves exploring the placenta with a fetoscope to localize the problematic vessel connections (anastomoses). These connections are then photocoagulated with a laser. The limited field of view of the fetoscope leads to poor spatial orientation during surgery, which makes it difficult for the surgeon to correctly identify the anastomoses. To address this problem, creating a 2D mosaic of the placenta has been proposed previously by [10, 27]. This technique expands the limited field of view of the fetoscope and hence augments scene available to the surgeon.

Standard mosaicking algorithms [4] use a projective transformation to model the relation between images assuming planarity or quasi-planarity in the imaged object. Subsequently, all images are propagated to a common plane on the basis of the computed transformations, forming a mosaic. Consecutive transformations are estimated in a pairwise fashion, which leads to an accumulation of error. This error gradually grows with the number of processed frames. More importantly, if one of the transformations fails to be estimated or suffers from a large degeneration, the mosaic cannot be computed.

The contributions of this paper are twofold. We introduce temporal information by using a Square Root Unscented Kalman Filter (SRUKF) to obtain a more robust mosaic. In addition, we use an external electromagnetic (EM) tracking system in combination with visual data that reduces the accumulation of error and further improves the robustness of the algorithm.

This paper is structured as follows: In section 2 we review the related work on mosaicking as well as tracking applied to image mosaicking. In section 3 we detail our algorithm. In section 4 we present our results obtained with a synthetic, phantom as well as an ex vivo dataset. We discuss various aspects of the algorithm in section 5 and draw conclusions and comment on future work in section 6.

2. Related work

Mosaicking has been used in many applications in the literature such as geographical 2D map reconstruction from aerial vehicles [7, 8], panoramas [4], among many others [14, 25]. The simplest approaches estimate a projective transformation or homography between successive images, thus assuming planarity in the imaged object. The use of feature-based approaches such as SIFT/SURF [2, 20] to obtain a transformation from corresponding interest points has become a standard procedure to generate 2D mosaics. These have the advantage of being more robust against non-uniform illumination than intensity-based approaches such as [1]. Nonetheless, the effectiveness of this technique in fetoscopy becomes compromised by the low quality of the interest points and the number of false correspondences that bypass standard outlier removal techniques such as RANSAC [16]. In [27], Reeff et al. proposed introducing a heuristic after RANSAC that imposes boundaries in the quality of the estimated homography by restricting the range of the determinant as well as imposing a minimum in the number of keypoints. They also proposed an algorithm to detect and discard mismatches.

A second challenge is the accumulation of error between successive frames, which becomes significant as the number of iterations increase. This is due to the pairwise fashion in which the mosaic is composed. In [4], a 2D bundle alignment was proposed to obtain a globally consistent mosaic using the correspondences between all images. Vercauteren et al. [14] explored a combination of rigid and deformable approaches, tackling the problem of global alignment by iteratively adding new pairwise rigid results to estimate the global parameters in a clinical environment. In [11], the global alignment was applied in clinical context as well. The idea of detecting a crossover i.e. the path of the camera returning to a previously imaged position, with the purpose of compensating the drift is exploited in [8], whereas in [22] a sequential bundle adjustment is performed by augmenting the state vector of a Variable State Dimension Filter (VSDF). Such a filter takes advantage of the diagonal structure of the covariance matrix to reduce the complexity of the algorithm. Even though the accumulation of error can be eliminated using these strategies, they are computationally very expensive.

Other approaches suggest employing an external tracking device to provide a global reference and reduce the accumulation of error. In [15], Yang et al. use a static 3D ultrasound probe to estimate the pose of the camera and build a mosaic using a combination of three methods: direct homography estimation, pose tracking and pose estimation from the ultrasound image. In [8], the use of the Global Positioning System (GPS) allows Unmanned Aerial Vehicles (UAV) to build a drift-free mosaic. Caballero et al. describe an on-line mosaicking technique using the Extended Kalman Filter (EKF), which takes advantage of the framework in order to include the GPS data by using the nRT decomposition of the homography. Our method is inspired by this technique; however, we aim to provide a reduced drift mosaic without the need of locating the crossover. In [6], an Extended Iterated Kalman Filter framework is generalized for when the observation and process models evolve in Lie groups. Mountney et al. [23] use SLAM with an EKF where the visible features form the state vector in order to provide a 3D approximation of the extended view that can be used as a navigational aid.

3. Methods

When the camera is imaging a planar object, the acquired images are related by a homograpy $\mathbf{H}_k$. Given that we just consider pairwise homographies, only the sub-index of the current time instant is kept for conciseness.

Using a pinhole camera model for a pre-calibrated fetoscope with $\mathbf{K}$ as the intrinsic matrix, the N corresponding points between frames are denoted by $\left\{\mathbf{p}_{k-1}^i, \mathbf{p}_k^i\right\}_{i=1}^N$ at time $k-1$ and k respectively. These points are related in an ideal noise-free scenario through the following equation:

$$\lambda \begin{bmatrix} \mathbf{p}_k^i \\ 1 \end{bmatrix} = \lambda \mathbf{q}_k^i = \mathbf{K}\mathbf{H}_k\mathbf{K}^{-1}\mathbf{q}_{k-1}^i \tag{1}$$

Where $\mathbf{q}_k^i$ is a point in homogeneous coordinates, $\mathbf{p}_k^i$ is a point in Cartesian coordinates and λ is the scalar associated to the homogeneous coordinates.

We compute these correspondences using SIFT, apply RANSAC to remove outliers and estimate a homography by using the well established DLT [26] algorithm.

3.1. Theoretical background

We introduce the generic dynamic state-space models framework to highlight the need for temporal and measurement equations. Since the information provided by the EM tracker is a 3D rigid motion transformation, it is more convenient to parameterize the state vector with the rotation and translation of the camera as well as the information of the imaging plane. For this purpose, the nRT decomposition is also introduced in this section.

3.1.1 Dynamic state-space models

The purpose of dynamic state-space models is to estimate the current world state given the observations from all time instants. Let us define the set of noisy measurements $\{\mathbf{z}\}_{i=1}^{N}$ that come from a set of world state variables $\{\mathbf{x}\}_{i=1}^{N}$. The world state estimates of points in a frame are not independent from the ones in the past frames; therefore, by using the Bayes rule, the probability of the state vector given all measurements can be expressed as:

$$Pr(\mathbf{x}_k|\mathbf{z}_{k,..,1}) = \frac{Pr(\mathbf{z}_k|\mathbf{x}_k)Pr(\mathbf{x}_k|\mathbf{z}_{k-1,..,1})}{\int Pr(\mathbf{z}_k|\mathbf{x}_k)Pr(\mathbf{x}_k|\mathbf{z}_{k-1,..,1})\mathbf{d}\mathbf{x}_k} \quad (2)$$

The first element of the numerator in equation 2, $Pr(\mathbf{z}_k|\mathbf{x}_k)$ corresponds to the measurement model, which defines the relation between the noisy measurement and the world state vector. The second element of the numerator can be expressed as the well-known Chapman-Kolmogorov relation [26]. By making the Markovian assumption, the current state depends only on the last state.

$$Pr(\mathbf{x}_k|\mathbf{z}_{k-1,..,1}) = \int Pr(\mathbf{x}_k|\mathbf{x}_{k-1})Pr(\mathbf{x}_{k-1}|\mathbf{z}_{k-1,..,1})\mathbf{d}\mathbf{x}_{k-1} \quad (3)$$

$Pr(\mathbf{x}_k|\mathbf{x}_{k-1})$ is the temporal model which specifies a temporal relation between adjacent time instants. $Pr(\mathbf{x}_{k-1}|\mathbf{z}_{k-1,..,1})$ is the posterior probability of the last iteration. Therefore, in order to model the probability of the state vector given the measurements of past time instants, a temporal and a measurement model must be defined.

3.1.2 The nRT Decomposition

In the case where the 3D object corresponds to a plane, a homography models the relation between corresponding points in two images. This homography can be decomposed into a rotation matrix $\mathbf{R}_k$, translation vector $\mathbf{t}_k$, the distance d_{k-1} from the optical center $\mathbf{O}_{k-1}$ of the camera at time $k-1$ to the plane and the normal vector $\mathbf{n}_{k-1}$ seen from the reference frame of the first camera [21, 24] as shown in figure 1.

$$\mathbf{H}_k = \mathbf{R}_k + \frac{\mathbf{t}_k}{d_{k-1}}\mathbf{n}_{k-1}^T \quad (4)$$

3.2. Our model

By modeling the relation between interest points in adjacent frames as a homography, the method can be used in quasi-planar environments, which are the real target scenarios. Making use of the nRT decomposition, we define the state vector as follows.

The state vector This encodes the rotation and translation between consecutive frames as well as the normal vector to the plane. We define $\mathbf{v}_{k-1}$ as the unit vector $\mathbf{n}_{k-1}$ divided by the distance d_{k-1}. Since d_{k-1} is not needed any further, it is not included as an extra parameter to estimate.

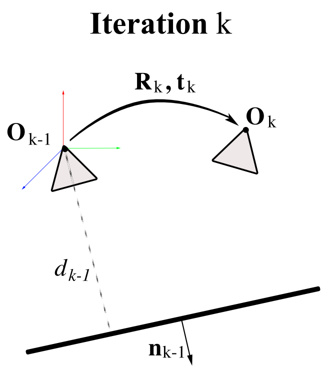

Figure 1: Two consecutive camera positions at time $k-1$ and k are imaging a plane. The homography relating the points in both images can be decomposed as a set which describe rotation $\mathbf{R}_k$, translation $\mathbf{t}_k$ and normal to the plane $\mathbf{n}_{k-1}$, divided by the distance d_{k-1} from the plane to the optical center of the camera at time $k-1$. The optical centers of the two cameras are denoted $\mathbf{O}_{k-1}$ and $\mathbf{O}_k$ respectively.

$$\mathbf{x}_k = \begin{bmatrix}\mathbf{r}_k^T & \mathbf{t}_k^T & \mathbf{v}_{k-1}^T\end{bmatrix}^T \quad (5)$$

where $\mathbf{r}_k, \mathbf{t}_k$ and $\mathbf{v}_{k-1}$ are respectively:

$$\mathbf{r}_k = \begin{bmatrix}r_k^x & r_k^y & r_k^z\end{bmatrix}^T \quad (6)$$

$$\mathbf{t}_k = \begin{bmatrix}t_k^x & t_k^y & t_k^z\end{bmatrix}^T \quad (7)$$

$$\mathbf{v}_{k-1} = \begin{bmatrix}\frac{n_{k-1}^x}{d_{k-1}} & \frac{n_{k-1}^y}{d_{k-1}} & \frac{n_{k-1}^z}{d_{k-1}}\end{bmatrix}^T \quad (8)$$

From the components of the rotation vector $\mathbf{r}_k$ in the state vector, the rotation matrix $\mathbf{R}_k$ in the special orthogonal group $\mathbb{SO}(3)$ [3] is obtained by using the Lie matrix exponential as in equation 9. The advantage of this parametrization is that orthogonality is directly imposed in the estimation of the rotation parameters.

$$\mathbf{R}_k = \exp\begin{pmatrix}0 & -r_k^z & r_k^y \\ r_k^z & 0 & -r_k^x \\ -r_k^y & r_k^x & 0\end{pmatrix} \quad (9)$$

It should be noted that parameterizing $\mathbf{r}$ in this form also implies a non-linear nature in the estimation of the rotation.

Even though the homography has eight degrees of freedom, nine parameters are used. This is because the plane is also encoded in the state vector.

The temporal model When the motion does not vary rapidly between frames (as in our fetoscopic video sequences), the rotation and translation can be modeled with a Brownian motion which corresponds to a constant velocity of the fetoscope, as in equation 10 and 11.

$$\mathbf{r}_k = \mathbf{r}_{k-1} + \epsilon_k^{p,r} \text{ with } \epsilon_k^{p,r} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}^{p,r}) \tag{10}$$

$$\mathbf{t}_k = \mathbf{t}_{k-1} + \epsilon_k^{p,t} \text{ with } \epsilon_k^{p,t} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}^{p,t}) \tag{11}$$

The noise terms $\epsilon_k^{p,r}$ and $\epsilon_k^{p,t}$ are modeled as Gaussian random variables with zero mean and covariance matrix $\mathbf{\Sigma}^{p,r}$ and $\mathbf{\Sigma}^{p,t}$ respectively. The temporal evolution of the normal obeys the following equation.

$$\mathbf{v}_{k-1} = \frac{\mathbf{R}_{k-1}\mathbf{v}_{k-2}}{1 + \mathbf{v}_{k-2}^T \mathbf{t}_{k-1}} + \epsilon_k^{p,v} \text{ with } \epsilon_k^{p,v} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}^{p,v}) \tag{12}$$

The noise $\epsilon^{p,v}$ is allowed in the evolution of the normal to account for slight deviations in the planarity assumption. The super-index p indicates that it is part of the temporal model.

The proof of equation 12 is presented here. The vector $\mathbf{v}_k$ is related to $\mathbf{n}_k$ and d_k as follows.

$$\mathbf{v}_k = \frac{\mathbf{n}_k}{d_k} \tag{13}$$

Firstly, given that we are observing a plane and assuming that the plane does not move, a translation does not change the direction of the normal vector. Therefore, the relation between $\mathbf{n}_{k-1}$ and $\mathbf{n}_k$ is:

$$\mathbf{n}_k = \mathbf{R}_k \mathbf{n}_{k-1} \tag{14}$$

The scalar d_k is the distance between the optical center $\mathbf{O}_k$ and the plane. The vector from the origin of coordinates to $\mathbf{O}_k$ corresponds to the translation vector $\mathbf{t}_k$.

$$d_k = \frac{\mathbf{n}_{k-1}^T \mathbf{t}_k + d_{k-1}}{|\mathbf{n}|} = \mathbf{n}_{k-1}^T \mathbf{t}_k + d_{k-1} \tag{15}$$

Finally,

$$\mathbf{v}_k = \frac{\mathbf{n}_k}{d_k} = \frac{\mathbf{R}_k \mathbf{n}_{k-1}}{d_{k-1} + \mathbf{n}_{k-1}^T \mathbf{t}_k} = \frac{\mathbf{R}_k \mathbf{v}_{k-1}}{1 + \mathbf{v}_{k-1}^T \mathbf{t}_k} \tag{16}$$

The measurement model This gives the relation between corresponding points in adjacent frames. For simplicity, we only model the noise in $\mathbf{p}_k^i$ and treat $\mathbf{p}_{k-1}^i$ as given.

$$\hat{\mathbf{q}}_k^i = \lambda \mathbf{q}_k^i = \mathbf{K}(\mathbf{R}_k + \mathbf{t}_k \mathbf{v}_{k-1}^T)\mathbf{K}^{-1} \begin{bmatrix} \mathbf{p}_{k-1}^i \\ 1 \end{bmatrix} \tag{17}$$

$$\mathbf{p}_k^i = \frac{\hat{\mathbf{q}}_{k,1:2}^i}{\hat{q}_{k,3}^i} + \epsilon_k^{i,m} \tag{18}$$

Where $\mathbf{p}_k^i$, the point in Cartesian coordinates at time k, is modeled as Gaussian random variable $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma}^m)$. The super-index m indicates that these entities are part of the measurement model.

The rotation and translation between adjacent frames computed from the global information provided by the EM tracker allow us to constrain the system. These relate to the state vector by:

$$\mathbf{r}_k^{EM} = \mathbf{r}_k + \epsilon_k^{EM,r} \text{ with } \epsilon_k^{EM,r} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}^{EM,r}) \tag{19}$$

$$\mathbf{t}_k^{EM} = \mathbf{t}_k + \epsilon_k^{EM,t} \text{ with } \epsilon_k^{EM,t} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}^{EM,t}) \tag{20}$$

3.2.1 The Square Root Unscented Kalman Filter (SRUKF)

This is a derivative-free, non-linear state and parameter estimation technique where the square root of the covariance matrix $\mathbf{S}$ is sampled in a set of the so called sigma points and then propagated. It is shown in [19] that it consistently outperforms the EKF in prediction and estimation. If the set of sigma points are chosen adequately, the algorithm can be accurate to the 3rd order term of the Taylor series for Gaussian inputs, and to the 2rd order term for non-Gaussian inputs. Two non-linearities are presented in our scenario: the temporal model for the normal vector in equation 12 and the measurement model for the correspondences in equation 17. The SRUKF [28] uses the Unscented Transform [18] to solve non-linear problems. It consists of a deterministic sampling of the input distribution in the so-called sigma points, which are later propagated through the non-linear function. Finally, a Gaussian distribution is approximated from the points as weighted mean and covariance. A set of $2L + 1$ sigma points is chosen (L is the length of the state vector). The choice of the sigma points as well as the weights can be optimized in order to minimize the error of the true non-linear function with respect to the modeled distribution. We refer the reader to [17] for more information about the optimality of the choice of the sigma points. Our choice of sigma points $\mathbb{X}^*$ is the following.

$$\mathbb{X}^* = [\hat{\mathbf{x}} \quad \hat{\mathbf{x}} + \gamma\mathbf{S} \quad \hat{\mathbf{x}} - \gamma\mathbf{S}] \tag{21}$$

where $\hat{\mathbf{x}}$ is the central point, corresponding to the zeroth weight, γ is defined as $\gamma = \sqrt{L + \iota}$. The weights of each sigma point j for the mean and covariance are denoted respectively with the super-indices μ and Σ:

$$w_0^{\mu} = \frac{\iota}{L + \iota} \tag{22}$$

$$w_0^{\Sigma} = \frac{\iota}{L + \iota} + (1 - \alpha^2 + \beta) \tag{23}$$

$$w_j^{\mu} = w_j^{\Sigma} = \frac{1}{2(L+\iota)} \qquad i = 1, ..., 2L \tag{24}$$

where $\iota = \alpha^2(L+\kappa) - L$, α controls the spread of the sigma points and it is usually set between 10^{-3} and 1, κ is a secondary scaling parameter usually set to 0, L is the length of the state vector and β is used to take advantage of the distribution if it is known *a priori*. For Gaussian distributions, the optimal value of β is 2 [30].

The most computationally expensive operation in the Unscented Kalman Filter (UKF) is the square root of the covariance matrix, which is usually performed as a Cholesky decomposition. The SRUKF tackles this problem by directly propagating the square root of the covariance matrix leading to a gain in efficiency from $\mathcal{O}(L^3)$ in the general UKF to $\mathcal{O}(L^2)$ where L is the number of dimensions of the state vector. In addition, by propagating the square root of the covariance matrix, symmetry and positive semi-definiteness are guaranteed. Since the w_0^{Σ} can be negative, it needs to be updated separately as explained in [28].

4. Results

The setup used to perform the experiments consists of a laparoscope Viking 3DHD[1] as well as the NDI Aurora system with a planar field generator and a Mini 6DOF sensor.[2] The setup is shown in figure 2. The data were obtained using only one channel of the laparoscope to simulate a monocular fetoscope. The synchronized video and EM tracking data was using the NifTK [9] software. Camera intrinsic and hand-eye calibration was performed using a 3 mm checkerboard, also implemented in the NifTK and described in [13]. Even though the image quality of the laparoscope is slightly better than in the fetoscope, the evaluation of the proposed algorithm is presented to be used for fetoscopy. In addition, we used the Matlab framework VLFeat [29] as basis for the implemented algorithms.

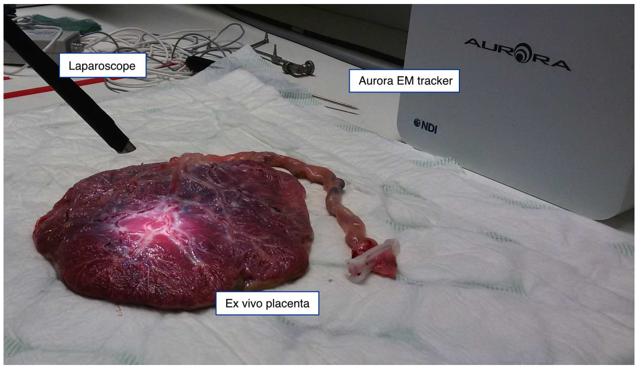

Figure 2: Using the laparosope and the Aurora EM tracker in an ex vivo placenta.

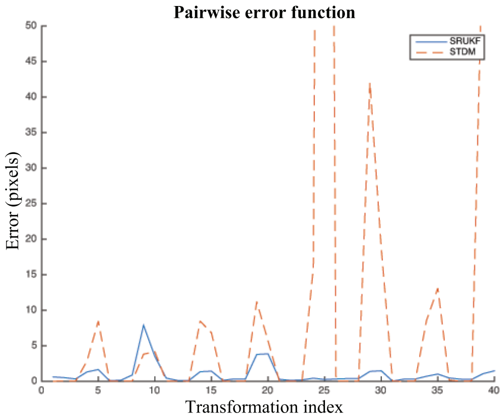

Figure 3: While the STDM shows high error peaks in the SYN dataset, the SRUKF manages to overcome them by using prior knowledge.

We created three datasets: a synthetic (SYN), a phantom (PHA) and an ex vivo (EXP) dataset. The SYN dataset was created from an image and a collection of homographies. We extracted a sequence of images by applying the homographies to a region of interest in the center of the image. Therefore, we ensure that the motion of the generated dataset obeys exactly a homographic motion. The PHA dataset consists of a handheld spiral scan of a printed image of a placenta. Even though the dataset is still far from clinical data, it allows us to test our algorithm when the assumption of planarity is fulfilled. The EXP dataset was created following the same motion pattern by scanning a real placenta. The main challenges of the latter are the reduction in quality of the interest points and the fact that the planarity assumption is not longer fulfilled, even though the scene can be considered quasi-planar.

We compare our model (SRUKF) against two algorithms: the standard pairwise mosaicking pipeline as is described in [4] (STDM) and 2D bundle adjustment (BA), the reference algorithm for reduction of accumulation of error. The comparison criteria between two homographies is the following. We project a grid of points with each homography and compute the mean of the Euclidean distance of the residual difference.

The datasets have been carefully designed to assess two main points: First, the robustness of our system to incorrect correspondences compared to standard algorithms [20]. Second, the potential improvement in accumulation of error. Our approach works in a sequential manner, achieving a substantial gain in computational efficiency while obtaining similar results to the BA. Our choice of values of covariance matrices is provided in the appendix and further commented in the discussion section.

Since the final mosaic relies on the pairwise composition of all frames, if no temporal information is used and

[1] http://www.conmed.com

[2] http://www.ndigital.com/medical/products/aurora

the data association is wrong, the composition will not be performed correctly. A high peak of error in a pairwise homography will bias the entire mosaic towards a wrong direction, and all the subsequent registrations will not be globally well aligned. In our model, the temporal evolution is used to produce a smoothing effect and avoid undesirable behaviors. As first experiment, we simulate a specific situation in which not enough quality interest points are obtained by adding peaks of noise to the images every five frames in the SYN dataset. Figure 3 presents in a quantitative way how the SRUKF manages to smooth the spikes of error resulting in an error reduction.

To demonstrate the achieved reduction in the accumulation of error in the PHA dataset, the mosaic is built using STDM, SRUKF and BA. Figure 4a shows the misregistration of a vessel in different frames (100 frames apart) due to the accumulation of error using the STDM. Figure 4b shows the resulting mosaic using our method with multiband blending. The accumulation of error is corrected successfully showing little difference to the BA in figure 4c. The reference image is shown in figure 4d for visual comparison.

To further provide quantitative results on the experiment, we have compared all homographies from the reference to each time instant for all algorithms. We obtained an increasing error tendency for the STDM as expected. This can be clearly seen in figure 5a. In figure 5b, any spike in the pairwise error results in an increase of the accumulation of error. In figure 5c, the mean trajectory of the grid of points is shown, comparing it with the mean trajectories for STDM and BA.

Lastly, we provide qualitative results in the EXP dataset (Figure 6), where the assumption of planarity is violated. While STDM is not able to cover the entire area of the placenta, our algorithm successfully creates full 2D map of the area.

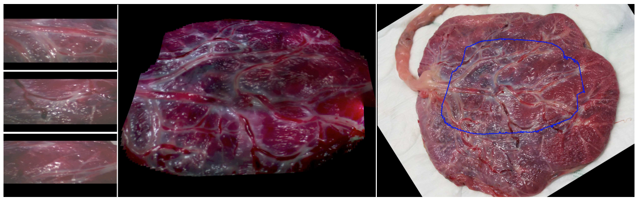

Figure 6: On the left, sample input images. On the middle, the mosaic of the EXP dataset using the SRUKF. On the right, the original image. The blue line indicates where the mosaic has been performed.

5. Discussion

In probabilistic temporal models, the temporal information is introduced in the form of a prior (Equation 2). If its covariance matrix decreases, i.e. the system relies more in the prior information, the estimation will be biased towards the prior knowledge. Otherwise, the estimation will tend to be just a maximum likelihood estimation. Therefore, there exists a trade-off between the temporal and measurement model. If the temporal model is right, then we can give it more weight, e.g. in the case of the fetoscope moving, we assume a constant velocity model of the fetoscope. If it is the case, then the temporal model will positively contribute to the estimation. Nonetheless, if there is a sudden twist in the motion of the fetoscope, the difference between the measurement and the prediction (so called innovation [26]) will grow and the temporal information will mislead the data.

In our case, we treat all covariance matrices as diagonal, i.e. all the variables are independent. In the measurement model, we have information about the relation between interest points in the images (Equation 17) as well as information from the EM tracker (Equations 19 and 20). Depending on the relation between their covariance matrices, either the EM tracking data or the interest points become more important.

On the one hand, if the EM tracker dominates the estimation, the system becomes more robust against accumulation of error. As we impose just the rotation and translation but not the normal, the system is constrained. On the other hand, when interest points drive the estimation, a more accurate homography is obtained. On the contrary, there is accumulation of error. The right choice of the covariance matrices then lies in a balance between interest points and EM tracking data, as well as temporal and measurement models. If the temporal covariance matrix is too small, the system will not have enough freedom to reach the right estimation, whereas if it is too large, the temporal model will have an adverse effect on the estimation.

6. Conclusions

We introduce a probabilistic temporal model that improves the robustness of the system by applying a strong temporal prior. In addition, we tackle the problem of the accumulation of error by incorporating global tracking data from an external EM tracking system by means of the nRT decomposition. We demonstrate qualitatively and quantitatively that our approach produces more robust and globally consistent mosaics than the STDM.

The limitations of the algorithm are (i) the assumption of planarity and (ii) the features. As future work, the covariance matrices must be learned from the data. In addition, the model can be upgraded to be piecewise-planar. On the other hand, the analysis of different types of features can provide more accuracy by strengthening the data association.

Further improvement in the model involves the use of the EM tracking data parametrized with absolute rather than

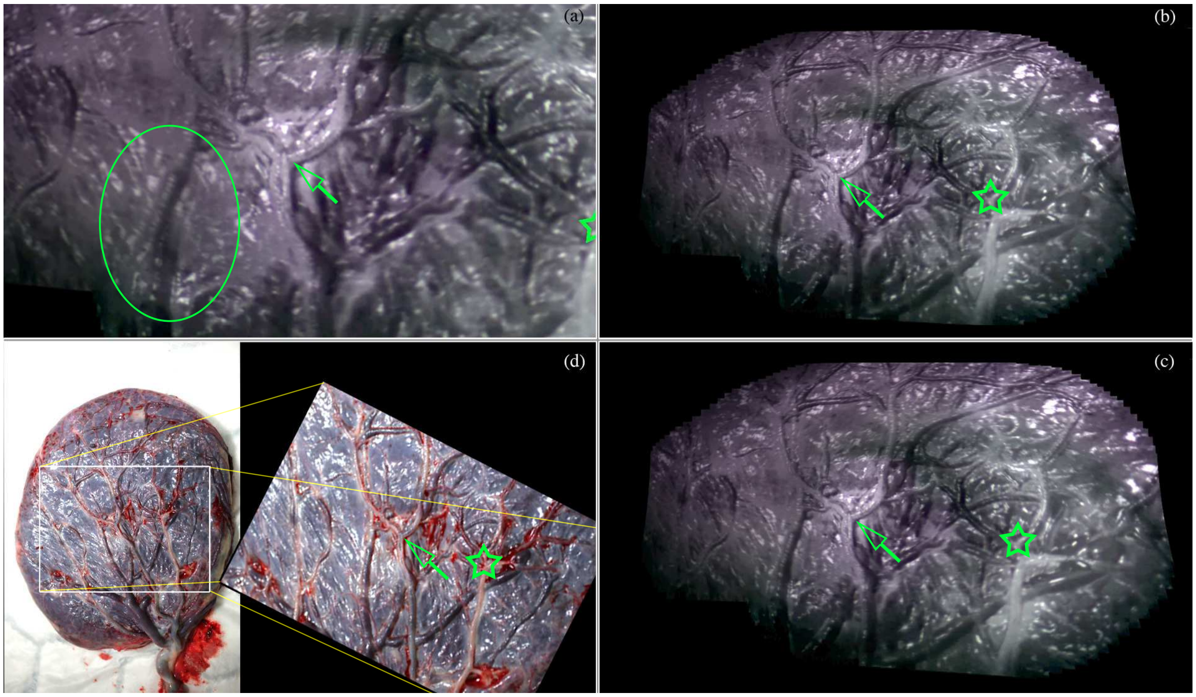

Figure 4: Visual effects of the accumulation of error in the PHA dataset. The green arrow and star are visual aims in order to facilitate the identification of the vessels to the reader. The results show that: (a) STDM. The vessel marked with a green oval is missaligned. This is due to the accumulation of error. (b) SRUKF. (c) BA. (d) Original image and a zoomed and rotated version to facilitate the visualization.

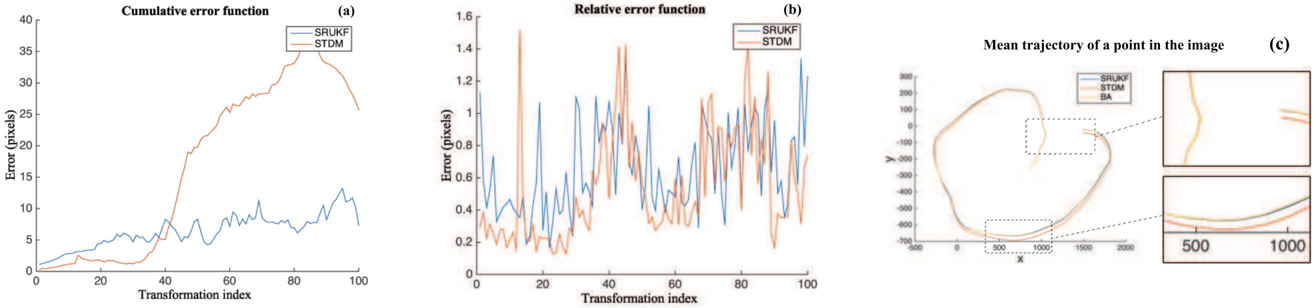

Figure 5: Quantitative results in the PHA dataset. (a) The cumulative function shows the tendency of the accumulation of error. The STDM increases whereas the SRUKF remains approximately constant. (b) The pairwise error. Any peak contributes to a large drift in the final mosaic for all the latter images. (c) The mean trajectory of a grid of points in the image is shown for STDM, SRUKF and BA. A clear drift from the STDM can be seen in the zoomed regions.

relative transformations. This will eliminate completely the accumulation of error, allowing for indefinitely long mosaics.

Appendix: Covariance matrix choices

The values of the covariance matrices have been chosen empirically. We use the term *diag* to refer to a matrix where all the values except the diagonal are zero.

$$\mathbf{\Sigma}^{p,r} = diag([1 \times 10^{-4}, 6.6 \times 10^{-6}, 1 \times 10^{-4}])$$

$$\mathbf{\Sigma}^{p,t} = diag([1.4, 1.05, 0.22])$$

$$\mathbf{\Sigma}^{p,n} = diag([1 \times 10^{-10}, 1 \times 10^{-10}, 1 \times 10^{-10}])$$

$$\mathbf{\Sigma}^{EM,r} = diag([1 \times 10^{-7}, 1 \times 10^{-7}, 1 \times 10^{-7}])$$

$$\mathbf{\Sigma}^{EM,t} = diag([1 \times 10^{-3}, 1.16 \times 10^{-3}, 0.23 \times 10^{-3}])$$

$$\mathbf{\Sigma}^{m} = 0.5 \times diag([1, 1, 1])$$

$$\mathbf{x}_0 = [0,0,0,0,0,0,0,0,0.02]^T, \mathbf{\Sigma}_0 = 10 \times \mathbf{\Sigma}^p$$

where $\mathbf{\Sigma}^p$ is the diagonal block matrix having $\mathbf{\Sigma}^{p,r}$, $\mathbf{\Sigma}^{p,t}$ and $\mathbf{\Sigma}^{p,n}$ as components.

Acknowledgements

This work was supported through 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 Fonds voor Wetenschappelijk Onderzoek Vlaanderen(FWO; JD as clinical researcher 1.8.012.07). 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). Sebastien Ourselin receives funding from EPSRC (EP/H046410/1, EP/J020990/1, EP/K005278) and the MRC (MR/J01107X/1). Marcel Tella and George Dwyer are supported by the EPSRC-funded UCL Centre for Doctoral Training in Medical Imaging (EP/L016478/1). We would also like to thank Efthymios Maneas for the contribution of the ex vivo placenta.

References

[1] Simon Baker and Iain Matthews. Lucas-Kanade 20 Years On : A Unifying Framework : Part 1 2 Background : Lucas-Kanade, 2004.

[2] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool. Speeded-Up Robust Features (SURF). *Comput. Vis. Image Underst.*, 110(3):346–359, 2008.

[3] Selim Benhimane and Ezio Malis. Homography-based 2D visual servoing. In *Proc. - IEEE Int. Conf. Robot. Autom.*, volume 2006, pages 2397–2402, 2006.

[4] Matthew Brown and David G Lowe. Recognising panoramas. *Proc. IEEE Int. Conf. Comput. Vis.*, pages 1218–1225, 2003.

[5] Baschat A et al. Twin-to-twin transfusion syndrome (TTTS). *J. Perinat. Med.*, 39(2):107–112, 2011.

[6] Bourmaud, Guillaume et al. From Intrinsic Optimization to Iterated Extended Kalman Filtering on Lie Groups. *J. Math. Imaging Vis.*, pages 1–20, 2016.

[7] Caballero, F. et al. Homography based kalman filter for mosaic building. Applications to UAV position estimation. In *Proc. - IEEE Int. Conf. Robot. Autom.*, pages 2004–2009, 2007.

[8] Caballero, F. et al. Unmanned Aerial Vehicle Localization Based on Monocular Vision and Online Mosaicking. *J. Intell. Robot. Syst.*, 55:323–343, 2009.

[9] Clarkson M et al. The NifTK software platform for image-guided interventions: platform overview and NiftyLink messaging. *Int. J. Comput. Assist. Radiol. Surg.*, 10(3):301–316, 2015.

[10] Daga, Pankaj et al. Real-time mosaicing of fetoscopic videos using sift. Feb 2016.

[11] Seshamani, Sharmishtaa et al. Direct global adjustment methods for endoscopic mosaicking. *SPIE Med. Imaging*, pages 72611D—-72611D, 2009.

[12] Slaghekke, Femke et al. Fetoscopic laser coagulation of the vascular equator versus selective coagulation for twin-to-twin transfusion syndrome: An open-label randomised controlled trial. *Lancet*, 383(9935):2144–2151, 2014.

[13] Thompson, Stephen et al. Hand–eye calibration for rigid laparoscopes using an invariant point. *International Journal of Computer Assisted Radiology and Surgery*, pages 1–10, 2016.

[14] Vercauteren, Tom et al. Mosaicing of confocal microscopic in vivo soft tissue video sequences. In *Lect. Notes Comput. Sci.*, volume 3749 LNCS, pages 753–760, 2005.

[15] Yang, Liangjing et al. Self-contained image mapping of placental vasculature in 3D ultrasound-guided fetoscopy, 2015.

[16] Martin a. Fischler and Robert C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24(6):381–395, 1981.

[17] Simon J. Julier and Jeffrey K. Uhlmann. Unscented filtering and nonlinear estimation. In *Proc. IEEE*, number 3, pages 401–422, 2004.

[18] S.J. Julier. The scaled unscented transformation. *Proc. 2002 Am. Control Conf. (IEEE Cat. No.CH37301)*, 6(2):4555–4559, 2002.

[19] Sj Julier and Jk Uhlmann. A New Extension of the Kalman Filter to Nonlinear Systems. *Int Symp AerospaceDefense Sens. Simul Control.*, 3(2):26, 1997.

[20] David G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.*, 60(2):91–110, 2004.

[21] Ezio Malis and Manuel Vargas. Deeper understanding of the homography decomposition for vision-based control. *Sophia*, 6303(6303):90, 2007.

[22] Philip F. McLauchlan and Allan Jaenicke. Image mosaicing using sequential bundle adjustment. In *Image Vis. Comput.*, number 9-10, pages 751–759, 2002.

[23] Peter Mountney and Guang-Zhong Yang. Dynamic view expansion for minimally invasive surgery using simultaneous localization and mapping. In *Engineering in Medicine and Biology Society, 2009. EMBC 2009. Annual International Conference of the IEEE*, pages 1184–1187. IEEE, 2009.

[24] H. Opower. Multiple view geometry in computer vision. *Opt. Lasers Eng.*, 37:85–86, 2002.

[25] Oscar Pizarro and Hanumant Singh. Toward large-area mosaicing for underwater scientific applications. *IEEE J. Ocean. Eng.*, 28(4):651–672, 2003.

[26] Simon Prince. Computer Vision (Models, Learning, and Inference) Algorithms. In *Comput. Vis. (Models, Learn. Inference)*, pages 1–75. 2013.

[27] Mireille Reeff, Friederike Gerhard, and Philippe Cattin. Mosaicing of Endoscopic Placenta Images. *GI Jahrestagung*, 93(1):467–474, 2006.

[28] Ronell Van Der Merwe and Eric a. Wan. The square-root unscented Kalman filter for state and parameter-estimation. *Acoust. Speech, Signal Process. 2001. Proceedings. (ICASSP -01). 2001 IEEE Int. Conf.*, 6:3461–3464, 2001.

[29] Andrea Vedaldi and Brian Fulkerson. VLFeat - An open and portable library of computer vision algorithms. *Design*, 3(1):1–4, 2010.

[30] Eric A Wan and Ronell Van Der Merwe. The unscented kalman filter for nonlinear estimation. pages 153–158, 2000.