ORIGINAL ARTICLE

CrossMark

Retrieval and registration of long-range overlapping frames for scalable mosaicking of in vivo fetoscopy

Loïc Peter[1] · Marcel Tella-Amo[1] · Dzhoshkun Ismail Shakir[1] · George Attilakos[3] · Ruwan Wimalasundera[3] · Jan Deprest[1,2] · Sébastien Ourselin[1] · Tom Vercauteren[1,2]

Abstract

Purpose The standard clinical treatment of Twin-to-Twin transfusion syndrome consists in the photo-coagulation of undesired anastomoses located on the placenta which are responsible to a blood transfer between the two twins. While being the standard of care procedure, fetoscopy suffers from a limited field-of-view of the placenta resulting in missed anastomoses. To facilitate the task of the clinician, building a global map of the placenta providing a larger overview of the vascular network is highly desired.

Methods To overcome the challenging visual conditions inherent to in vivo sequences (low contrast, obstructions or presence of artifacts, among others), we propose the following contributions: (1) robust pairwise registration is achieved by aligning the orientation of the image gradients, and (2) difficulties regarding long-range consistency (e.g. due to the presence of outliers) is tackled via a bag-of-word strategy, which identifies overlapping frames of the sequence to be registered regardless of their respective location in time.

Results In addition to visual difficulties, in vivo sequences are characterised by the intrinsic absence of gold standard. We present mosaics motivating qualitatively our methodological choices and demonstrating their promising aspect. We also demonstrate semi-quantitatively, via visual inspection of registration results, the efficacy of our registration approach in comparison with two standard baselines.

Conclusion This paper proposes the first approach for the construction of mosaics of placenta in in vivo fetoscopy sequences. Robustness to visual challenges during registration and long-range temporal consistency are proposed, offering first positive results on in vivo data for which standard mosaicking techniques are not applicable.

Keywords Fetoscopy · Image mosaicking · Image registration

Introduction

Twin-to-twin transfusion syndrome (TTTS) is a condition affecting identical twin pregnancies, where unexpected vascular anastomoses occur between two twins sharing a single placenta [3]. This results in a blood imbalance between the two twins. The current state-of-the-art curative procedure consists in laser photo-coagulation via fetoscopy of the abnormal vessel anastomoses located on the placenta. More precisely, surgeons perform a progressive visual exploration of the placenta, with the aim of localising and eliminating the anastomoses which allow a direct blood transfer between the two twins. Due to the difficulty of manipulating the fetoscope and due to the very limited field-of-view available at each timepoint to the surgeon, some anastomoses can be missed by the surgeon leading to an only incomplete treatment [12]. To assist a clinician during TTTS surgery, mosaicking approaches are desirable to create a map of the placenta from a video acquired during fetoscopy. With the help of such a map, the field-of-view can be enlarged to facil-

✉ Loïc Peter
  l.peter@ucl.ac.uk

[1] Wellcome/EPSRC Centre for Interventional and Surgical Sciences, University College London, London, UK

[2] Department of Development and Regeneration, Cluster Woman and Child, Centre for Surgical Technologies, KU Leuven, Leuven, Belgium

[3] University College Hospital, London, UK

itate the task of the clinician regarding the identification of yet unexplored areas of the placenta, and to provide a better overview of the topology of the vascular network.

Image mosaicking is a classical computer vision problem, where the panorama of a scene is built from a series of overlapping pictures. The most standard approach for stitching images consists of the registration of overlapping pairs via the detection and matching of landmarks [5]. Such feature-based methods have been successfully applied for some medical applications, such as retinal mosaicking [15] and fibroscopic video mosaicking [1]. However, other type of clinical images may display a lack of texture, occlusions and other factors that make a landmark-based registration of a pair of images not reliable enough. To address this issue, alternative registration methods have been employed such as semi-dense registration method for dynamic view expansion in an in vivo porcine experiment [21]. In [13], dense correspondences in an in vivo experiment were used, demonstrating improvements over RANSAC-based algorithms. Another example of dense registration was also successfully applied for confocal microscopy [23]. We refer to [4] for a more comprehensive review of the intersection between simultaneous localisation and mapping (SLAM), scene reconstruction and mosaicking in endoscopic procedures.

Closer to our application case, some works have attempted to perform mosaicking in placental images. In [16], the authors report challenging situations that they tackle using a modified RANSAC algorithm. In [9], a robust matching in phantom data was proposed via a new feature extractor algorithm using CNNs. External modalities such as 3D ultrasound [11] or an electromagnetic tracker [20] were also investigated as means of guiding the mosaicking process. However, these approaches addressing the mosaicking of placenta images were until now limited to phantom and ex vivo data, for which visual properties are considerably different from the in vivo cases encountered in clinical scenarios. In the latter conditions, challenges such as repeated occlusions (e.g from foetal limbs or impurities present in the amniotic fluid, see Fig. 1) and low image contrast do not allow a successful application of standard landmark-based computer vision techniques.

In this paper, we propose a first approach towards the generation of placental mosaics from in vivo fetoscopy data. Our method combines (1) a registration method based on the alignment of gradient orientations, ensuring robustness to visual challenges inherent to in vivo acquisitions, and (2) a strategy based on bags of visual words which allows the identification of pairs of overlapping frames located at arbitrary time points in the sequence. By retrieving and registering these key pairs of frames, the global consistency of the mosaic can be improved in a scalable manner. Qualitative results are reported and discussed based on real sequences and demonstrate first promising results towards the clinical

use of mosaicking methods for TTTS surgery. In addition, we inspected visually the results of pairwise registration on an example sequence and labelled manually their quality, showing the benefit of our approach in comparison with two standard baselines: registration based on the robust matching of SURF-based keypoints, and dense image alignment based on normalised cross-correlation.

An approach for in vivo mosaicking

Problem statement

Each frame of a fetoscopy sequence offers a partial view of the imaged placenta. Under the assumption that the placenta is planar, two arbitrary frames of the sequence are related by a homography transformation. Formally, given a sequence of N frames $I_1, \ldots, I_N$ where each $I_i : \Omega_i \rightarrow \mathbb{R}^3$ is an RGB image defined over a domain $\Omega_i \subset \mathbb{R}^2$, there exists for every pair of images (I_i, I_j) a homographic warping $w_{j,i} : \Omega_i \rightarrow \Omega_j$ such that for every $\mathbf{x} \in \Omega_i$, $I_i(\mathbf{x})$ and $I_j(w_{j,i}(\mathbf{x}))$ are visual measurements corresponding to the same location of the placenta. To create and visualise a mosaic, we propose the following approach. First, without loss of generality, the first frame of the sequence is defined as a reference frame located within the central part $\Omega_{\text{abs}}^1 \subset \Omega_{\text{abs}}$ of a (sufficiently large) mosaic image $M : \Omega_{\text{abs}} \rightarrow \mathbb{R}^3$. The mosaicking task aims at stitching together overlapping images to create a global map of the placenta or, in other words, at warping and placing each frame of the sequence on the corresponding part of the mosaic domain Ω_{abs}. For every frame I_i, the corresponding subset Ω_{abs}^i of the absolute mosaic domain Ω_{abs} must be found. Equivalently, we propose to estimate a global homography W_i such that $\Omega_{\text{abs}}^i = W_i(\Omega_{\text{abs}}^1)$. Note that W_1 is already defined as the identity via our choice of reference frame. Moreover, global and relative warpings are related via $W_i \circ W_j^{-1} = w_{i,j}$. To estimate the global warpings, we rely primarily on a series of pairwise registrations of overlapping frames which are directly conducted in the relative image domains Ω_i.

A fully sequential approach for mosaicking would register all consecutive frames and, with the obtained relative warpings $w_{k,k-1}$ for $2 \leq k \leq i$, compute the global warping W_i of the i-th frame as follows:

$$W_i = w_{i,i-1} \circ w_{i-1,i-2} \circ \cdots \circ w_{2,1}. \tag{1}$$

If the estimation of the relative warpings is perfect, the equation above allows in theory a perfect mosaicking. However, in practice, errors in the estimation of each relative warping accumulate so that a clear mismatch can be observed when the fetoscope comes back to a previously visited location. This effect can even degenerate if the presence

Fig. 1 Visual challenges in in vivo fetoscopy. Nearly consecutive frames of an in vivo sequence are shown. Together with a low contrast, in vivo data are subject to more or less severe occlusions due to foetal limbs or impurities present in the amniotic fluid
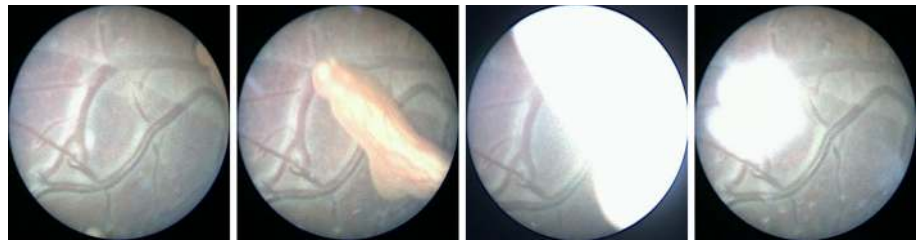
of occlusions makes the pairwise registration of two consecutive frames unfeasible, thereby breaking the chain of transformations (1). For increased robustness and temporal consistency over a large number of frames, it is therefore desired to register additional overlapping frames that are not necessarily consecutive (for example, the frames obtained when revisiting a portion of the placenta). Overall, if we denote $\mathcal{R}$ the set of couples of indices (i, j) for which a registration has been performed and for which a resulting (possibly noisy) warping $\hat{w}_{i,j}$ estimating the true warping $w_{i,j}$ has thus been computed, and noticing that $W_i \circ W_j^{-1} = w_{i,j}$, we can look for global warpings $\hat{W}_1, \ldots, \hat{W}_N$ such that

$$\hat{W}_1, \ldots, \hat{W}_N = \operatorname*{argmin}_{W_1,\ldots,W_N} \sum_{(i,j)\in\mathcal{R}} d(W_i \circ W_j^{-1}, \hat{w}_{i,j}), \qquad (2)$$

where d is a measure of dissimilarity between two warpings. The formulation (2) is closely related to bundle adjustment and was proposed by Vercauteren et al. in the context of rigid transformations [23]. We define the distance $d(w_1, w_2)$ between two warpings w_1 and w_2 defined over a rectangular image domain Ω as follows. First, we decide on a discrete set of reference points $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_m\} \subset \Omega$, which we choose in our case as a regular grid of step 3 over Ω. The distance $d(w_1, w_2)$ is then defined as

$$d(w_1, w_2) = \max_{\mathbf{x}\in\mathcal{X}} \|w_1(\mathbf{x}) - w_2(\mathbf{x})\|_2. \qquad (3)$$

This allows us to obtain an intuitive geometrical interpretation of the distance between warpings as the maximum deviation in terms of Euclidean distance over the set of reference points $\mathcal{X}$.

After having found estimates of the absolute warpings by solving (2), a final mosaic can be created with blending algorithms [6,14]. In this paper, we focus on the accurate assessment of the global warpings, i.e. on the correct placement of the frames of the sequence on the mosaic. We used a standard publicly available technique [6] to generate the mosaics shown in this paper (Fig. 4).

To summarise, we identified two crucial components for mosaicking:

1. Given two overlapping frames I_i and I_j, we need a robust and reasonably fast way to register them to obtain a warping $w_{i,j}$.

2. To improve the consistency of the estimation over long timeframes, it is crucial to identify additional overlapping frames that are not consecutive but located at timepoints arbitrarily far from another.

We propose in this work a strategy to address separately these two challenging problems, respectively, exposed in "Pairwise registration of consecutive frames" and "Ensuring long-range consistency with bag of words" section, which takes into account the visual properties of in vivo sequences.

Pairwise registration of consecutive frames

The traditional image stitching technique [5] based on the detection and matching of landmarks (e.g. with a combination of SIFT and RANSAC) is prone to failure in in vivo sequences encountered in clinical conditions. The lack of constrast in the acquired images and the cluttered and varying aspect of the observed scene are challenges responsible for these difficulties. In this section, we present a registration method that addresses the pairwise registration of in vivo frames. Given the aforementioned challenges, we propose not to rely on landmarks. Instead, we perform a dense pixelwise alignment of the gradient orientations and propose a variant of the maximisation of the correlation of image gradients introduced by Tzimiropoulos et al. [22], with two main differences exposed in details below. Aligning gradient orientations possesses several advantages: it is for example invariant to local changes of contrast and is suitable for registering accurately linear structures such as vessels, which matches the main clinical objective of mosaicking for TTTS surgery, i.e. the creation of an overview of the topology of the placental vascular network. Moreover, by focusing solely on gradient orientations and not on the gradient norms, each pixel is given the same weight, which naturally improves the robustness of the registration to visual artifacts and partial occlusions.

The registration task consists in estimating the true warping $w_{j,i}$ such that $I_i(\mathbf{x})$ and $I_j(w_{j,i}(\mathbf{x}))$ correspond to the same location for every $\mathbf{x} \in \Omega_i$. With this formulation, I_i is called the fixed image and I_j the moving image. We parametrise the homographic warpings with a vector $\mathbf{p} \in \mathbb{R}^8$ corresponding to the 8 coefficients of the canonical homographic representation, i.e. such that

$$w((x, y), \mathbf{p}) = \left(\frac{p_1 x + p_2 y + p_3}{p_7 x + p_8 y + 1}, \frac{p_4 x + p_5 y + p_6}{p_7 x + p_8 y + 1} \right).$$

$$(4)$$

As discussed above, we propose to look for the registration warping $\hat{w}_{j,i}$ which aligns best the gradients of I_i and I_j. Since we explicitly do not want to take into account the strength of the gradients, we first normalise the gradients of the fixed and moving images ensuring a unit gradient norm at every pixel. For a point $\mathbf{x} \in \Omega_i$ of the domain of the fixed image, we denote $\Delta\theta(\mathbf{x}, \mathbf{p})$ the angle between the gradient of the fixed image and the gradient of the warped moving image at $\mathbf{x}$. We define the final warping $\hat{w}_{j,i}$, i.e. the output of our registration method, as $\hat{w}_{j,i} = w(., \hat{\mathbf{p}})$ where

$$\hat{\mathbf{p}} = \underset{\mathbf{p}}{\operatorname{argmin}} \sum_{\mathbf{x} \in \Omega_i} \sin^2 \Delta\theta(\mathbf{x}, \mathbf{p}). \qquad (5)$$

Since $\sin^2 t = \frac{1}{2}(1 - \cos 2t)$, the proposed approach can be seen as a variant of the maximisation of the correlation of image gradients [22] defined by

$$\hat{\mathbf{p}} = \underset{\mathbf{p}}{\operatorname{argmax}} \sum_{\mathbf{x} \in \Omega_i} \cos \Delta\theta(\mathbf{x}, \mathbf{p}). \qquad (6)$$

We can identify two main differences between the two formulations. First, our pixelwise costs based on the sine function are minimal for $\Delta\theta = 0$ or $\Delta\theta = \pi$, whereas the terms in (6) are minimal for $\Delta\theta = 0$ only. Thereby, only the orientation (modulo π) of the gradients is taken into account in our cost function, and not their direction. It appears to be a useful property in practice: as we try to match vessels with an iterative method (see below), optimisation steps must be able to cross areas where gradients are oriented in opposite directions before reaching the optimal vessel alignment. Having written our minimisation problem (5) as a sum of squares, we are also able to use known results on nonlinear least squares, and more precisely the forward additive version of the Lucas Kanade algorithm [2]. This formulation not only leads to simpler theoretical mathematical derivations, but also offers the possibility to use off-the-shelf optimised solvers for nonlinear least squares problem, such as the Ceres solver [17] which includes classical optimisation techniques (the Gauss–Newton, Levenberg–Marquardt or Powell Dog–Leg algorithms, for example).

Solving (5) with the Gauss–Newton algorithm To solve numerically the minimisation problem (5), we use the fact that it is a nonlinear least squares problem to apply the Gauss–Newton algorithm, which, in the context of image registration, can also be seen as the forward additive version of the Lucas–Kanade algorithm [2]. To keep the following derivations as general as possible, we denote $N = |\Omega_i|$ and arbitrarily order the elements of Ω_i so that $\Omega_i =$

$\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$, and we denote M the number of parameters encoding the transformation, i.e. the size of the parameter vectors $\mathbf{p}$. Applying the Gauss–Newton algorithm, we approximate iteratively the desired minimum with a series of parameters $\mathbf{p}^{(1)}, \mathbf{p}^{(2)}, \ldots$ such that

$$\mathbf{p}^{(k+1)} = \mathbf{p}^{(k)} - (\mathbf{J}^T \mathbf{J})^{-1} \mathbf{J}^T \mathbf{s}(\mathbf{p}^{(k)}), \qquad (7)$$

where $\mathbf{s}(\mathbf{p}^{(k)})$ is the $N \times 1$ column vector $(\sin \Delta\theta(\mathbf{x}_i, \mathbf{p}^{(k)}))_{1 \le i \le N}$ and $\mathbf{J}$ is the $N \times M$ Jacobian matrix whose coefficients J_{ij} are defined as

$$J_{ij} = \frac{\partial s_i(\mathbf{p}^{(k)})}{\partial p_j} = \frac{\partial \Delta\theta(\mathbf{x}_i, \mathbf{p}^{(k)})}{\partial p_j} \cos \Delta\theta(\mathbf{x}_i, \mathbf{p}^{(k)}). \qquad (8)$$

We denote $\mathbf{g}_m(\mathbf{x}_i, \mathbf{p}^{(k)}) = (g_{m,x}(\mathbf{x}_i, \mathbf{p}^{(k)}), g_{m,y}(\mathbf{x}_i, \mathbf{p}^{(k)}))$ the gradient of the warped moving image at the location $\mathbf{x}_i$, and $\theta_m(\mathbf{x}_i, \mathbf{p}^{(k)})$ (respectively, $\theta_f(\mathbf{x}_i)$) the angle of the gradient of the moving image (respectively, the fixed image) at the location $\mathbf{x}_i$. By definition, we have $\Delta\theta(\mathbf{x}_i, \mathbf{p}^{(k)}) = \theta_m(\mathbf{x}_i, \mathbf{p}^{(k)}) - \theta_f(\mathbf{x}_i)$ and

$$\theta_m(\mathbf{x}_i, \mathbf{p}^{(k)}) = \arctan \frac{g_{m,y}(\mathbf{x}_i, \mathbf{p}^{(k)})}{g_{m,x}(\mathbf{x}_i, \mathbf{p}^{(k)})}, \qquad (9)$$

so that the coefficients of the Jacobian given in (8) can be written more explicitly as

$$J_{ij} = \frac{1}{\|\mathbf{g}_m\|^2} \left(g_{m,x} \frac{\partial g_{m,y}}{\partial p_j} - g_{m,y} \frac{\partial g_{m,x}}{\partial p_j} \right) \cos \Delta\theta, \qquad (10)$$

where the dependencies in $\mathbf{x}_i$ and $\mathbf{p}^{(k)}$ were omitted for readability. We finally mention that, although the gradients $\mathbf{g}_m(\mathbf{x}_i, \mathbf{p}^{(k)})$ could be computed at each iteration by warping the moving image and computing the gradient of the resulting warped image numerically, it is more efficient to precompute once for all the gradient of the (unwarped) moving image and obtain $\mathbf{g}_m$ by warping this gradient and multiplying it by the Jacobian of the warping, by application of the chain rule [22]. The partial derivatives of $g_{m,x}$ and $g_{m,y}$ are obtained similarly.

Ensuring long-range consistency with bag of words

As discussed in "Problem statement" section, evaluating the set of global homographies from a series of pairwise registrations of consecutive frames inevitably leads to an accumulation of registration errors. In the most extreme case, the chain of transformations can even be broken if a registration is not feasible at all, for example in the presence of a full occlusion. However, fetoscopy conditions naturally

lead to long sequences during which the surgeon follows vessels one by one, resulting in the presence of numerous overlapping areas in the sequence. Therefore, introducing additional constraints from the registration of non-temporally consecutive frames may provide the redundancy to compensate for the drift and the robustness to failed registrations of consecutive frames. If we could have a reliable way to decide from the registration result if the registration was successful (see "Assessing the validity of a registration" section), we could in theory register all image pairs to extract the highest amount of information. However, registering all image pairs is computationally intractable for long sequences and would probably introduce more redundancy than required. Therefore, we need an efficient way to predict, from their visual appearance, the pairs that are worth registering.

Following an idea introduced in computer vision [10], we adopt a strategy based on bags of visual words [7] to efficiently identify frames sharing a similar content without the need to register them first. We sample dense keypoints using the VGG descriptor [19] and perform a K-means clustering over the full video to obtain a vocabulary of K visual words. Each image is then described by a signature vector $\mathbf{v} \in \mathbb{N}^K$ encoding the frequency of each visual word in the image. The visual similarity between two images I_i and I_j is then computed as the cosine distance between the two associated signature vectors $\mathbf{u}$ and $\mathbf{v}$, i.e

$$s(I_i, I_j) = \frac{\sum_{k=1}^{K} \mathbf{u}_k \mathbf{v}_k}{\sqrt{\sum_{k=1}^{K} \mathbf{u}_k^2} \sqrt{\sum_{k=1}^{K} \mathbf{v}_k^2}}. \tag{11}$$

By computing this similarity measure for every pair of images in the videos, we obtain a similarity matrix on which the revisiting of previous locations is apparent (Fig. 2). The construction of this matrix is more scalable than attempting the registration of all pairs and, in fact, only requires approximate nearest neighbours for which algorithms in linear time (e.g. FLANN) are available. Figure 2 shows an example of similarity matrix. In this example video, the trajectory followed by the clinician is "star-shaped": every vessel is followed until its extremity, before following it back until the last intersection, usually at the coord insertion site. The timepoints and patterns corresponding to these "back and forth" trajectories are apparent on the similarity matrix as lines orthogonal to the diagonal.

To define the set of additional candidate registrations to be included in the bundle adjustment formulation (2), we simply rely on a threshold on the similarity above which the registration of a pair is tried. The choice of this threshold is mainly driven by computational considerations, as it is directly related to the number of attempted registrations
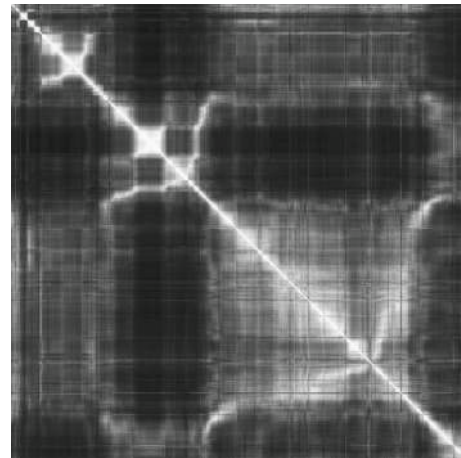

Fig. 2 Similarity matrix Every entry (i, j) of this matrix states how visually similar the frames I_i and I_j are. Note how the camera following a vessel back and forth creates branches that are orthogonal to the diagonal line

which are going to be performed before solving the bundle adjustment problem.

Assessing the validity of a registration

In the previous subsections, we described a robust method to register two images, as well as a way to retrieve pairs of non-consecutive images which share a visual overlap so that a registration of these pairs can be attempted, in addition to the consecutive pairs. Each registration is used as a term in the bundle adjustment formulation (2) and acts as a constraint in the estimation of the global homographies necessary to create the mosaic. However, in practice, the obtained registrations are not always accurate, due for example to the registration optimiser being trapped in a local minimum, to a failure of the retrieval of overlapping frames (leading to the unfeasible registration of two frames for which there is no overlap), or to the presence of a large occlusion in a frame caused for example by foetal limbs. Assessing the correct registrations within the attempted ones is of critical practical importance to filter out these wrong constraints added in the bundle adjustment, which would bias the final estimation of global homographies.

In this work, we declare a registration as successful if it is close enough from the identity (in the sense of the distance d defined in "Problem statement" section), and if the gradient-based cost function which was minimised in (5) is sufficiently small in comparison with the costs obtained with random warpings sampled around the identity. Although this empirical strategy proved to be effective in practice, we plan to investigate as future work more sophisticated methods, such as consistency checks over cycles of frames [8] or learning-based approaches.

Experimental validation

Implementation details

We implemented our method in C++ using the OpenCV library. The bundle adjustment minimisation problem (2) was solved numerically using the Levenberg–Marquardt algorithm with the Ceres solver [17]. By restricting the warpings to affine transformations, i.e. homographies where $p_7 = 0$ and $p_8 = 0$ with the notations introduced in (4), convergence to a visually sound solution was achieved, even when starting from identity warpings as initialisation. If one desired to work with general homographies instead, closed-form solutions could be used to obtain initialisations close enough from the global optimum [18]. The pairwise registrations using the forward additive Lucas Kanade algorithm were performed in a Gaussian pyramidal fashion with 6 levels (where each level is a blurred and scaled down version of the original image, as implemented in OpenCV), starting by registering the images at a coarse level and refining progressively the warpings by increasing the resolution. In the case where a registration is rejected at a level of the pyramid, it was reinitialised as the identity for the next level. For increased robustness, we performed for each pair both a forward and backward registration (i.e. switched fixed and moving image) and kept the registration leading to the lowest cost. The bags of visual words were computed using the OpenCV implementation, with default parameters in the extraction of the VGG descriptors.

The pairwise registration between two frames takes approximately 1 second. These registrations remain the main bottleneck in practice: on our example video of 600 frames, solving the bundle adjustment problem takes a few minutes, as does the construction of the similarity matrix from the bag of words. Therefore, there is a direct linear relationship between the total computational time and the amount of attempted registrations of retrieved pairs of frames. For a given computational budget, this time can be controlled by registering a predefined amount of pairs with the maximum similarity, as mentioned at the end of "Ensuring long-range consistency with bag of words" section. Given our current implementation and set of parameters, the total mosaicking pipeline took about 3 hours.

Qualitative results

Due to the nature of in vivo acquisitions, a ground truth for placenta mosaicking is not available. This makes a quantitative evaluation of our approach very difficult. Nevertheless, we demonstrate first promising visual results with our approach, which we discuss qualitatively in this section. Figure 3 shows, on an example video of 600 frames, the appearance of the mosaic obtained after the estimation of the global homography of each video frame. To facilitate the visual interpretation of the results and relate it more easily to the actual content of our example video, we limit the display of the mosaic to truncated versions of the video at different timepoints (from 1 to 4, chronologically). Each frame of the video is pasted chronologically onto the mosaic according to the global homography obtained after the offline optimisation. Note that the global optimisation has been run once for all on the whole video, and that these chronological timepoints are only introduced for the visualisation of our results. Figure 3 illustrates the global consistency of the mosaic: although an area showing a ≺-shaped vascular intersection is visited 4 times during the sequence (once at each chosen snapshot), this intersection is correctly placed at the same location in the mosaic over time. This is due to the fact that video frames containing this intersection have been recognised as similar in terms of visual content and successfully registered, adding additional constraints in the global optimisation which lead to improved temporal consistency.

We show in Fig. 4 the resulting mosaics, where frames are merged in a seamless fashion following the method of Burt and Adelson [6] publicly available in the software Enblend,[1] where one frame every 5 frames is used for blending. We show the mosaics obtained when our gradient-based registration method was used, with and without the long-range consistency with the bag-of-word formulation (respectively, Fig. 4a, b). In the purely sequential case (Fig. 4a), the drifting behaviour can be seen via the generally more distorted aspect of the mosaic, as well as clear misalignments (best seen when compared to Fig. 3), for example of the vessels on the top left part and of the membrane of the amniotic sac (linear demarcation on the bottom left part).

Manual evaluation of pairwise registrations

In addition to the qualitative results discussed in "Qualitative results" section, we compare our gradient-based registration approach with two baselines. Our first baseline is a classical image stitching technique used in computer vision available in OpenCV which consists in detecting SURF-based keypoints in each image and align the images with RANSAC. SURF features were chosen over SIFT as they performed slightly better empirically, confirming the observations made by Reeff [16] on other placenta images. Our second baseline is obtained by replacing our gradient-based similarity measure by the normalised cross-correlation, keeping the rest of our Gauss–Newton optimisation framework unchanged.

To conduct the comparison, we considered our example video of 600 frames and assessed visually the quality of the 599 sequential pairwise registrations for each baseline. Each

---

[1] http://enblend.sourceforge.net/.

Fig. 3 Retrieval and registration of overlapping frames for long-range consistency. By retrieving and successfully registering frames showing the same vascular intersection (marked with blue arrows), the location of this area in the mosaic is kept stable, ensuring an improved global consistency
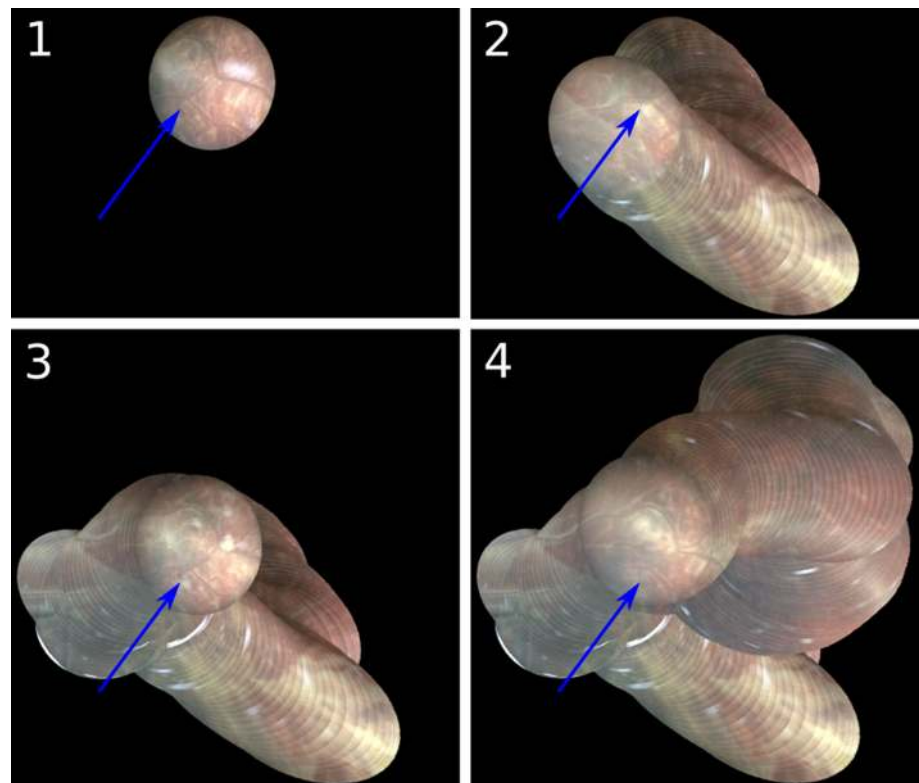

Fig. 4 Mosaics obtained after blending. We show two example mosaics obtained with and without the introduction of long-range consistency. Without long-range consistency (**a**), an accumulation of errors between pairwise registrations occurs, so that misalignments are caused when revisiting locations (such as the vessels on the top left part or the membrane on the bottom left part, both marked with blue arrows). **a** Purely sequential alignments. **b** Our approach
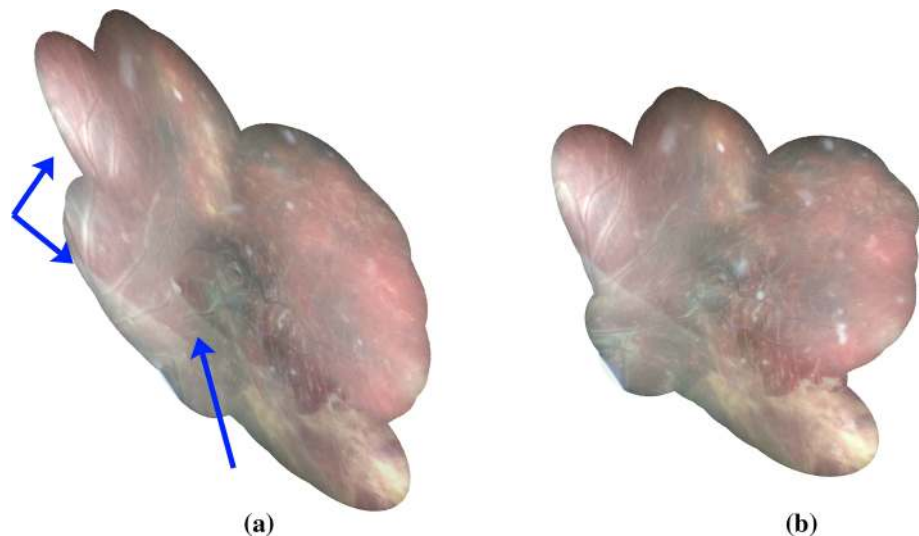
registration was manually labelled as either *correct*, *incorrect* or *doubtful* (for ambiguous cases where the correctness of the registration is difficult to assess visually). To remove any subjective bias, the 1797 registrations to be labelled were randomly shuffled so that the annotation was done without knowledge of the approach which was actually evaluated in each case. Table 1 summarises the count of registration belonging to each category for the three methods and confirms the benefit of our approach based on the alignment of gradient orientations.

Conclusion

We proposed a first step towards the mosaicking of placenta images from in vivo fetoscopy data. A robust registration method based on the alignment of gradient orientations addresses the visual challenges inherent to in vivo sequences which prevent the successful application of classical mosaicking techniques used in computer vision. Moreover, the global consistency of the mosaics was improved via a retrieval strategy based on bags of visual words, which

Table 1 Evaluation of pairwise registrations

Method	Correct registrations	Doubtful	Incorrect registrations
SURF−RANSAC	56 (9.3%)	83 (13.9%)	460 (76.8%)
NCC	29 (4.8%)	61 (10.2%)	509 (85.0%)
Ours	477 (79.6%)	67 (11.2%)	55 (9.2%)

identifies pairs of frames which are worth registering, regardless of their respective location in time. Qualitative results were shown and discussed to illustrate the relevance of our approach.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Ethical approval All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards.

Informed consent Informed consent was obtained from all individual participants included in the study.

References

1. Atasoy S, Noonan D, Benhimane S, Navab N, Yang G (2008) A global approach for automatic fibroscopic video mosaicing in minimally invasive diagnosis. In: MICCAI 2008. Springer, pp 850–857
2. Baker S, Matthews I (2004) Lucas–Kanade 20 years on: a unifying framework. Int J Comput Vis 56(3):221–255
3. Baschat A, Chmait R, Deprest J, Gratacós E, Hecher K, Kontopoulos E, Quintero R, Skupski D, Valsky D, Ville Y (2011) Twin-to-twin transfusion syndrome (TTTS). J Perinat Med 39(2):107–112
4. Bergen T, Wittenberg T (2016) Stitching and surface reconstruction from endoscopic image sequences: a review of applications and methods. IEEE J Biomed Health Inform 20(1):304–321
5. Brown M, Lowe D (2007) Automatic panoramic image stitching using invariant features. Int J Comput Vis 74(1):59–73
6. Burt P, Adelson E (1983) A multiresolution spline with application to image mosaics. ACM Trans Graph (TOG) 2(4):217–236
7. Csurka G, Dance C, Fan L, Willamowski J, Bray C (2004) Visual categorization with bags of keypoints. In: ECCV Workshop on statistical learning in computer vision, pp 1–22
8. Datteri R, Liu Y, D'Haese P, Dawant B (2015) Validation of a nonrigid registration error detection algorithm using clinical MRI brain data. IEEE Trans Med Imaging 34(1):86–96
9. Gaisser F, Jonker P, Chiba T (2016) Image registration for placenta reconstruction. In: IEEE conference on computer vision and pattern recognition workshops (CVPRW), pp 473–480
10. Ho K, Newman P (2007) Detecting loop closure with scene sequences. Int J Comput Vis 74(3):261–286
11. Liao H, Tsuzuki M, Kobayashi E, Dohi T, Chiba T, Mochizuki T, Sakuma I (2008) Fast image mapping of endoscopic image mosaics with three-dimensional ultrasound image for intrauterine treatment of twin-to-twin transfusion syndrome. In: Medical imaging and augmented reality: 4th international workshop. Springer, pp 329–338
12. Lopriore E, Middeldorp J, Oepkes D, Klumper F, Walther F, Vandenbussche F (2007) Residual anastomoses after fetoscopic laser surgery in twin-to-twin transfusion syndrome: frequency, associated risks and outcome. Placenta 28(2):204–208
13. Lourenço M, Stoyanov D, Barreto J (2014) Visual odometry in stereo endoscopy by using PEaRL to handle partial scene deformation. In: Linte C, Yaniv Z, Fallavollita P, Abolmaesumi P, Holmes D (eds) MICCAI workshop AE-CAI 2014. Springer, Cham, pp 33–40
14. Mahé J, Linard N, Tafreshi M, Vercauteren T, Ayache N, Lacombe F, Cuingnet R (2015) Motion-aware mosaicing for confocal laser endomicroscopy. In: MICCAI 2015. Springer, pp 447–454
15. Prokopetc K, Bartoli A (2017) SLIM (slit lamp image mosaicing): handling reflection artifacts. Int J Comput Assist Radiol Surg 12(6):911–920
16. Reeff M (2011) Mosaicing of endoscopic placenta images. Ph.D. Thesis
17. Sameer A, Keir M and others. Ceres solver. http://ceres-solver.org
18. Schroeder P, Bartoli A, Georgel P, Navab N (Jan 2011) Closed-form solutions to multiple-view homography estimation. In: IEEE workshop on applications of computer vision (WACV), pp 650–657
19. Simonyan K, Vedaldi A, Zisserman A (2014) Learning local feature descriptors using convex optimisation. IEEE Trans Pattern Anal Mach Intell 36(8):1573–1585
20. Tella M, Daga P, Chadebecq F, Thompson S, Shakir DI, Dwyer G, Wimalasundera R, Deprest J, Stoyanov D, Vercauteren T, Ourselin S (2016) A combined EM and visual tracking probabilistic model for robust mosaicking: application to fetoscopy. In: IEEE conference on computer vision and pattern recognition workshops (CVPRW), pp 524–532
21. Totz J, Mountney P, Stoyanov D, Yang G (2011) Dense surface reconstruction for enhanced navigation in MIS. In: Fichtinger G, Martel A, Peters T (eds), MICCAI 2011. Springer, pp 89–96
22. Tzimiropoulos G, Zafeiriou S, Pantic M (2011) Robust and efficient parametric face alignment. In: ICCV, pp 1847–1854
23. Vercauteren T, Perchant A, Malandain G, Pennec X, Ayache N (2006) Robust mosaicing with correction of motion distortions and tissue deformations for in vivo fibered microscopy. Med Image Anal 10(5):673–692