Journal of *Imaging*

MDPI

Article

Stable Image Registration for In-Vivo Fetoscopic Panorama Reconstruction †

Floris Gaisser [1,*], Suzanne H. P. Peeters [2], Boris A. J. Lenseigne [1], Pieter P. Jonker [1] and Dick Oepkes [2]

[1] BioMechanical Engineering, Delft University of Technology, 2628CD Delft, The Netherlands; B.A.J.Lenseigne@tudelft.nl (B.A.J.L.); P.P.Jonker@tudelft.nl (P.P.J.)
[2] Department of Obstetrics, Leiden University Medical Center, 2333ZA Leiden, The Netherlands; S.H.P.Peeters@lumc.nl (S.H.P.P.); d.oepkes@lumc.nl (D.O.)
* Correspondence: f.gaisser@tudelft.nl
† This paper is an extended version of our paper published in: Gaisser, F.; Peeters, S.H.; Lenseigne, B.; Jonker, P.P.; Oepkes, D. Fetoscopic Panorama Reconstruction: Moving from Ex-vivo to In-vivo. In Annual Conference on Medical Image Understanding and Analysis; Springer: Berlin, Germany, 2017; pp. 581–593.

Received: 31 October 2017; Accepted: 9 January 2018; Published: 19 January 2018

Abstract: A Twin-to-Twin Transfusion Syndrome (TTTS) is a condition that occurs in about 10% of pregnancies involving monochorionic twins. This complication can be treated with fetoscopic laser coagulation. The procedure could greatly benefit from panorama reconstruction to gain an overview of the placenta. In previous work we investigated which steps could improve the reconstruction performance for an in-vivo setting. In this work we improved this registration by proposing a stable region detection method as well as extracting matchable features based on a deep-learning approach. Finally, we extracted a measure for the image registration quality and the visibility condition. With experiments we show that the image registration performance is increased and more constant. Using these methods a system can be developed that supports the surgeon during the surgery, by giving feedback and providing a more complete overview of the placenta.

Keywords: panorama reconstruction; in-vivo fetoscopy; stable region detection

1. Introduction

The Twin-to-Twin Transfusion Syndrome (TTTS) is a condition involving monochorionic twins (twins with a shared placenta) with an imbalance in blood exchange that occurs through vascular anastomoses (connecting blood vessels) on the shared placenta. It occurs in about 10% of such pregnancies and is a condition that can lead to fatal complications for both twins [1]. It can be treated by fetoscopic laser surgery, a technique to separate the fetal circulation through fetoscopic laser coagulation. This surgery increases the survival rate over other techniques but relies on the condition that all vascular anastomoses have been found [2]. For this the surgeon has to scan the placenta for all places where blood vessels of the two twins connect and accurately note them. Then these vessels have to coagulated in a specific order to restore the blood exchange balance and finally separate the blood circulation of both twins. This scanning procedure is complicated due to the very limited field of view of the fetoscope, lack of proper landmarks on the placenta and bad visibility conditions. A panorama reconstruction of the placenta would greatly reduce the chance of complications of the surgery. Panorama reconstruction of internal anatomical structures has found many applications, such as for retina [3], bladder [4] and esophagus [5] reconstruction. However, fetoscopic panorama reconstruction has mostly been done ex-vivo [6–9].

For panorama reconstruction, it is necessary to correctly find all transformations between images constituting the panorama. A transformation between two adjacent images can be estimated by

matching key-points in both images, assuming they are correctly matched and the key-points accurately describe the same locations on the placenta. In our previous work we investigated what is necessary to move from ex-vivo to in-vivo fetoscopic panorama reconstruction and we identified specific challenges for an in-vivo setting: The visibility is complicated by the color and turbidity of the amniotic fluid. Also, the bounded motion of the fetoscope continuously changes the distance to the placenta. Lastly, the light intensity of the fetoscope is limited as one cannot blind the fetus. These aspects result in a very small range (Figure A1) in which current key-point methods for image registration can be fruitfully used.

To cope with the problems of an in-vivo setting, we suggest four points of improvement with respect to our previous work: Improve the key-point detection and matching method in order to achieve more robust image registration. Furthermore, detect unavailable or inaccurate image registrations and discard these from the panorama reconstruction. Also, not to create image registration chains, but to register to a part of the panorama. Finally, improve the visibility by obtaining an image quality measure and providing feedback to the surgeon to move the fetoscope in a certain way and the operation assistant to adjust the light source. Of course, also the equipment plays a role in the performance of the panorama reconstruction; i.e., a larger viewing angle improves the field of view and a high dynamic range or low light camera will obtain a larger range of feasible visibility conditions.

In this work in Section 2 we revisit the problems of in-vivo panorama reconstruction and we formulate requirements for proper reconstruction. Then in Section 3 we introduce recent developments in deep-learning and how this can be used to tackle the problems. In Section 4 we evaluate the proposed approach and in Section 5 we discuss the outcome and come to conclusions.

2. Challenges of In-Vivo Setting

In previous work we described key aspects in which an in-vivo setting differs from an ex-vivo setting and we concluded that in contrast with an ex-vivo setting, state-of-the-art key-point methods have a very limited performance in an in-vivo setting. Therefore other approaches e.g., based on deep learning must be found. In this section we recap the differences in setting and how they influence the image registration between two adjacent fetoscopic images, and we conclude with presenting a set of requirements for a proper image registration in in-vivo settings. The next section then describes the methods we propose to adhere to these requirements.

2.1. Differences in Setting

The visibility in fetoscopic images is a key problem that complicates the image registration between two or more images in an in-vivo setting. The first aspect of good visibility is the amount of light as well as an even distribution. In an ex-vivo setting, the amount of light can be completely controlled and positioned. Therefore, an optimal position and an even distribution of light can be obtained. However, in an in-vivo setting this is not the case:

- The amount of light is limited by the light source and cannot be chosen too bright as it might blind the fetus to the point of annoyance such that the fetus becomes restless.
- The amniotic fluid is far from clear as the fetus micurates in it. Moreover, as commonly the case in TTTS, the fetus might release bowel movements due to distress, giving the amniotic fluid a green turbid color. This color and turbidity of the amniotic fluid absorbs light, reducing the distance the light can reach.
- Also, the fetus and particles that float in the amniotic can limit the field of view of the view of the placenta. In Figures 1c and 2c air bubbles can be observed. However, these are the result of using water in mimicking the in-vivo setting and are not part of the surgical setting.
- The source of light is the fetoscope itself. This results in an uneven distribution of light, which reduces the amount of illumination towards the edge of the view. Furthermore, saturation of the imaging sensor in the center of the image inhibits proper observation of the structure of the placenta.

Examples are shown in Figure 1. Especially for the green turbid liquid it is difficult for the camera to acquire a proper image, resulting in a large amount of sensor noise.

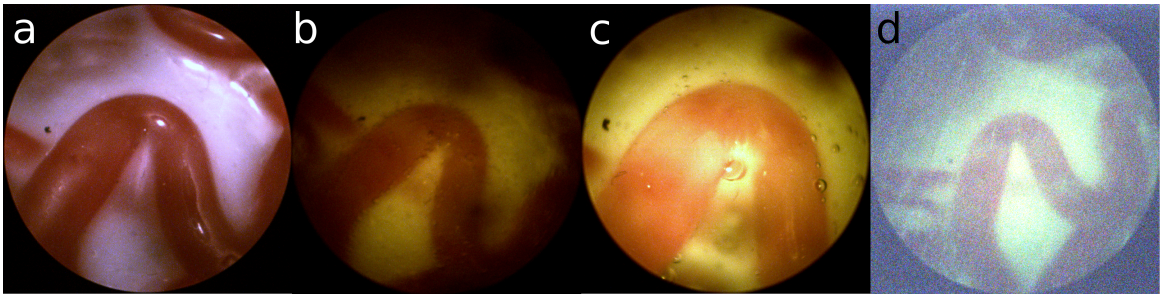

Figure 1. (**a**) Ex-vivo view; (**b**) uneven distribution of light; (**c**) too much light saturating the sensor; (**d**) not enough light creating sensor noise.

The second aspect of good visibility is the distance to the placenta. With enough distance to the placenta it is possible to observe many different structures on the placenta. In an ex-vivo setting the placenta is generally placed on a flat surface and the fetoscope can be positioned at any distance to the placenta. Furthermore, the fetoscope can be moved laterally with equal distance to the placenta. However, this is not the case in an in-vivo setting:

- The distance to the placenta is limited due to the reduced amount of light.
- The fetoscope is limited in motion at the point of entry. It can only rotate around the point of entry and move forward and backward.
- A lateral movement of the field of view can only be obtained by rotation. Therefore, the lateral change of view also changes the distance to the placenta. This results not only in a change of visible structure, but also a change in illumination.
- The scanning procedure in the in-vivo setting is to follow veins from the umbillical cord and back. Which creates large loops, whereas the ex-vivo setting uses a spiraling motion, which has many small loops.

Figure 2a shows an example of an ex-vivo setting with a satisfactory amount of structure. In contrast, Figure 2b shows a nominal example of an in-vivo setting with green turbid liquid. From this point of view, the fetoscope can be moved laterally resulting in a closer view (Figure 2c) or more distant view (Figure 2d). One can observe that suboptimal viewing conditions are unavoidable in an in-vivo setting.

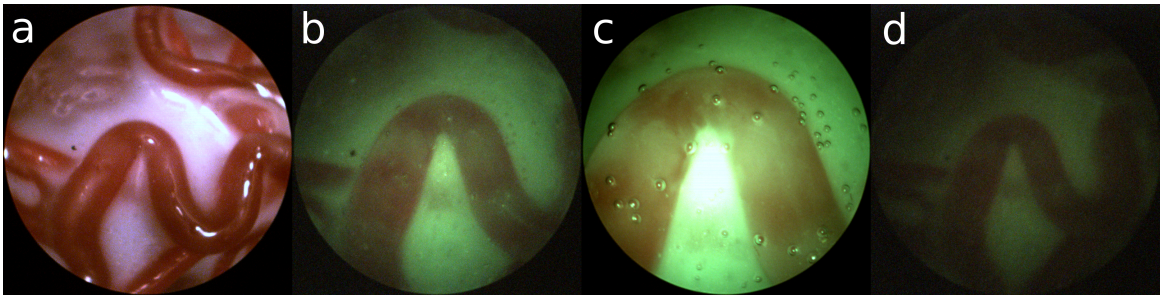

Figure 2. Ex-vivo: (**a**) sufficient structure; In-vivo: (**b**) nominal; (**c**) close and bright; (**d**) far and dark.

2.2. Influence on Image Registration

For panorama reconstruction, it is necessary to correctly find all transformations between adjacent images constituting the panorama. A transformation between two adjacent images can be estimated with a minimum of 4 matches, assuming they are correctly matched and accurately describe the

same locations on the placenta. The key-point matching process assumes that two well matching key-points describe the same physical point. To find matching key-points in two images, the area around a key-point is described with a histogram of gradients. Around corners this generally provides an unique enough description of the key-point such that it can be matched with a similar key-point in an adjacent image. Such a corner is dominated by equally strong gradients in two dimensions. In contrast, along edges, such as along a vein, there is a strong gradient perpendicular to the edge and practically no gradient along the edge. Consequently, key-points selected on an edge are very alike as they have a very similar structure around the point. Moreover, taking sensor noise into account, the histogram of gradients has an additional random component that is often larger than the fine difference between two edges in adjacent images. With a growing variation in the exact location and an increasing number of incorrect matches, the required number of correct matches increases as well. The *LMeDS* transform estimation method is robust to inaccurate locations, but requires at least 50% correct matches to obtain a transformation [10]. Whereas, the *RANSAC* method is sensitive to inaccurate locations, though robust to incorrect matches [11]. Unfortunately there is no method that is robust to both inaccurate locations and incorrect matches.

In an in-vivo setting the limited distance to the placenta reduces the observable structure and the limited amount of light creates sensor noise. Hence, unstable key-points are detected that are described by similar features and matching key-points result in many seemingly good, but incorrect matches, describing different points on the placenta, usually along veins. Concluding, in three key aspects traditional key-point matching methods fail in an in-vivo setting; detecting stable key-points, reliable matching of key-points, and obtaining enough matches for a proper estimation of the transform.

2.3. Image Registration Requirements

To research other approaches, such as based on deep learning, it is important to specify the requirements for an image registration process that consistently performs its task in an in-vivo setting:

- Key-points in one image should be reproducible in another image and both should accurately describe the same physical location on the placenta
- The features describing a key-point in one image should be so unique that the matching key-point in another image has almost the same unique features
- Key-points in one image for which no matching key-point is found in the other image should have such unique features that it is not incorrectly matched to key-points in that other image at different locations
- The image registration process should be able to detect whether an obtained transformation is incorrect in order to exclude it from the panorama reconstruction.

The section below describes the method we propose to adhere to these requirements.

3. Method

In recent years, deep-learning neural networks have been applied in many different fields, tackling various complex problems [12]. This approach is successful because it has the ability to learn any complex task without having knowledge on how to solve the task, as long as the desired output is known and enough training data is available. A deep learned network consists of a pipeline of trainable layers, which makes it possible to train the network to handle compound structures.

Convolutional layers are very suitable to extract relevant data from structured data such as images. It is comparable to convolutional filtering the image, but then with filter coefficients that are trained instead of coefficients determined by a user. A convolutional layer has a set of filters that is moved over the input image extracting relevant structures everywhere in the image. This can be applied in many different applications, notably in image classification [13].

In this work we propose a deep convolution neural network to tackle the challenges stated in the previous section. With it we will:

- Detect stable regions on the veins of the placenta
- Extract matchable features from these regions
- Learn a visibility and matchability measure of an image

These steps are detailed in the next sub-sections.

3.1. Stable Region Detector

Soon after the introduction of deep learning it was also applied to the detection of key-points [14–16]. These methods are similar to handcrafted methods such as *SIFT* and *ORB*, but have the advantage that the networks can be trained to select key-points that are more apt for matching and image registration. Although these networks are often trained with key-points detected by a handcrafted method, this is not very suitable for our case and another way of obtaining a key-point training set needs to be found.

Image registration requires the detection of stable key-points, but it is yet unclear what defines a stable key-point not being a corner. A straight edge (Figure 3a) constraints the key-point in one direction. This is also the case for a circular edge, when rotation is also taken into account, as shown in Figure 3b. However, key-points with the same curvature can be matched. On curved edges, having an additional change in scale, the matching becomes more unique, but not unique enough to do the job (Figure 3c). Therefore, any edge alone, albeit curved, cannot be considered a source of stable key-points. We need additional information to make the key-point unique.

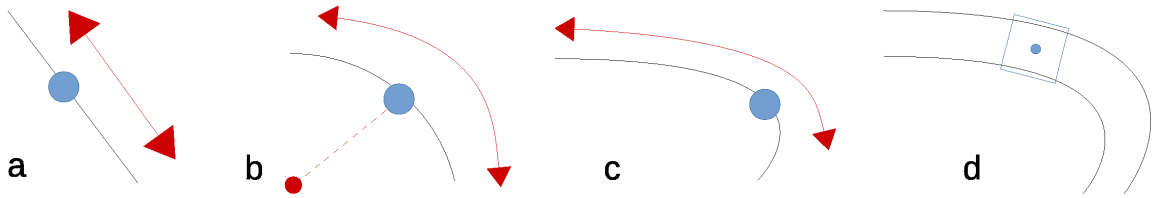

Figure 3. Constraints on (**a**) edge; (**b**) circular; (**c**) curve; (**d**) veins.

Consequently, we propose to define stable key-points being center points on the medial axis of the veins. As both sides of the vein are curves of different curvature they provide independent constraint dimensions making the point more unique. When also the width of the vein is taken into account this constraints the detection also in the dimension of scale, as shown in Figure 3d. This makes our proposed method less a key-point detector but rather a region detector; we use three instead of two independent dimensions.

Since our approach resembles region/object detection rather than key-point detection, we investigated also Region Convolutional Neural Networks such as RCNN [17], Fast-RCNN [18], Faster-RCNN [19] and Single Shot Detector (SSD) [20], which have been developed to detect and classify objects in images. Earlier methods such as RCNN and Fast-RCNN used external region proposal methods, but Faster-CNN and SSD use the same convolutional network for classification as for region proposals, where SSD detects objects at multiple scales. Therefore, this last method was chosen as basis for our stable region detection method.

The SSD method detects regions by defining bounding boxes with their min and max corners as shown in Figure 4a. These are learned by training the neural network to output the location of the two corners for each feature cell according to their *default boxes*. An additional classification layer learns the detection probability of each class in every default box. If the classification layer outputs a positive classification, the matching output of the detection layer is used for localization of the classified object. We refer to the Faster-RCNN [19] and SSD [20] papers for more details on the specifics on how to train these detectors.

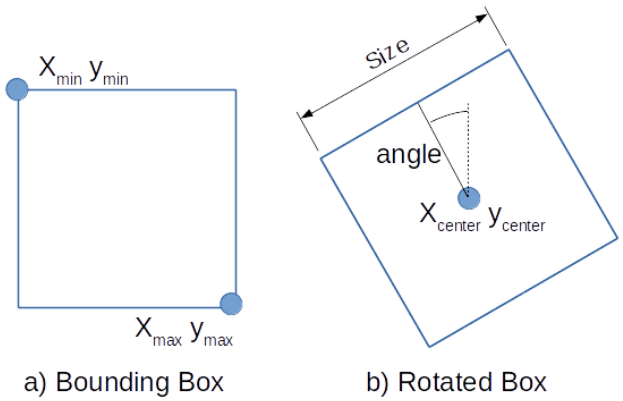

Figure 4. (**a**) Definition of Bounding Box (BBox); (**b**) Definition of Rotated Box (RBox).

In order to detect stable regions on the placenta, we propose to detect square areas on the veins. However, the bounding boxes as defined by SSD are not suitable to describe the orientation of the vein. Therefore, we extent SSD and redefine the default boxes by the center, the size, and the angle of the box, as shown in Figure 4b.

The ground truth of these detections is obtained by manually annotating the center and the radius of the veins in the images. Taking the gradient of these annotations, also the direction of the vein is defined. An example of such annotation of the veins is shown in Figure 5.

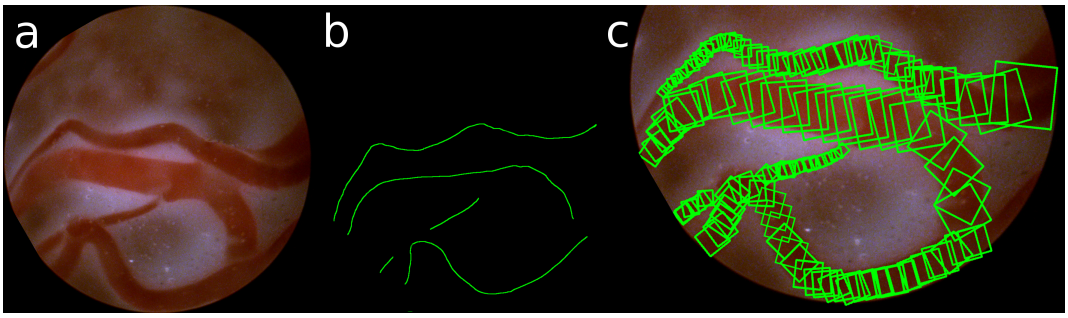

Figure 5. (**a**) Sample image; (**b**) annotated center line; (**c**) selection of annotated RBoxes.

It is interesting to note that the definition of our points of stable-regions are similar to that of key-points. Similar to a key-point we also extract features around a location of a rotated box. But whereas key-points are solely defined by a point, a scale and an orientation of the key-point, we restrict the possible locations to be at the center of a vein. This makes them stable and within some margin also reproducible.

3.2. Stable Matching

The second challenge in the image registration process is to extract features that are descriptive enough for the proper matching of key-points. In [15], this was achieved by training with positive and negative samples, using Euclidean distance to measure similarity in a Siamese CNN. This is similar to [7] where patches were selected in a grid to extract features that were trained in a Siamese CNN with contrastive loss.

In this paper we extended the SSD architecture similarly to [7]. An additional convolution layer extracts a feature for the detection of Contrastive Loss. Furthermore, every detection is fine-tuned with its matching performance such that detections that are difficult to match are assigned a lower probability to be detected. In this way we remain with matchable features.

3.3. Qualitative Measures

Our last challenge is to obtain a measure of success for the image registration. This can be used to guide the surgeon or/and his assistant. For this, a qualitative measure is trained by using the matching performance which was used to train matchable features. Since the images registration is highly influenced by the visibility, we define two more outputs to describe this visibility. One describes the amount of illumination and the other describes the distance to the placenta. The visibility is defined as optimal in nominal illumination and distance conditions.

These outputs provide an indication about the performance of the image registration. In case of bad registration the images can be discarded in the process. However, to obtain a sequence of images that is continuous, the surgical team should be included in the process, i.e., the surgeon should be made aware that the panorama reconstruction process has lost position. Furthermore, an assistant controlling the light intensity should be made aware of the illumination condition to actively adjust this.

3.4. Network Architecture

The above described contributions are implemented based on the VGG-16 network with SSD as a starting point. In order to detect stable regions, we first associate detection scales with the annotated veins of various sizes and select only the first four levels as a scale space pyramid to detect rotated boxes. Each detection scale by default consists of three layers; first the classification layer for determining if there is a positive detection. Second, the location layer describing the location of the detection and third the prior boxes, describing the template detections. Every scale also passes on the features to the next scale.

Next, for stable matching we change two aspects; First, the SSD network was made into two parallel pipelines as shown in Figure 6a. These two networks share their weights as a Siamese Neural Network. Second, each detection scale is extended with an additional convolutional layer to extract a feature describing every detection as shown in Figure 6b. These, combined with the region detections can be used to find the matches for image registration.

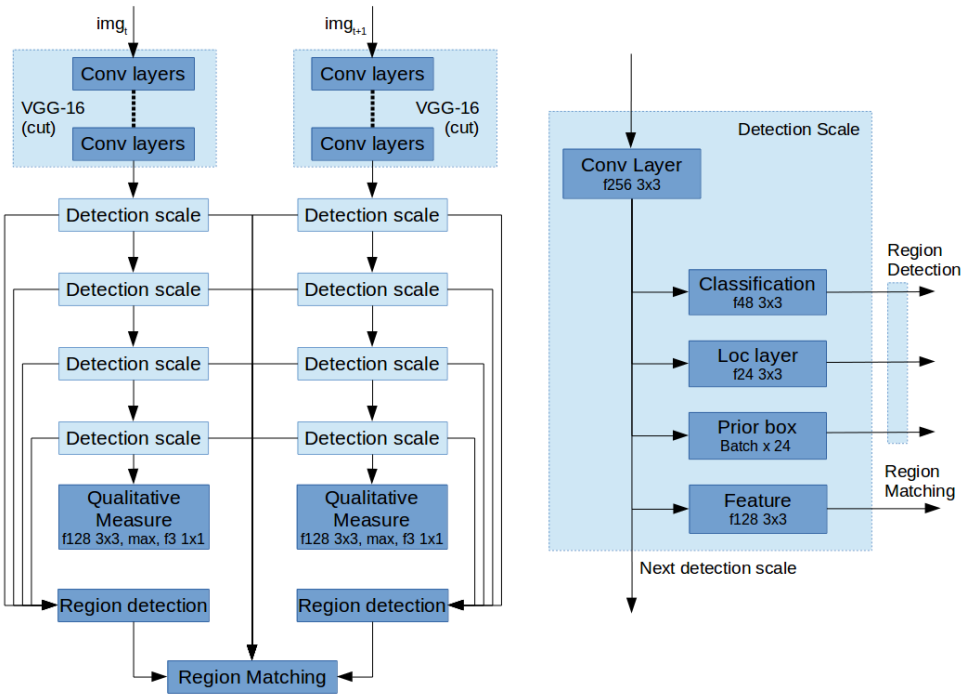

Figure 6. (**a**) Left: Detection network architecture; (**b**) Right: Architecture of a single detection scale.

Finally, to extract a measure for visibility and image registration performance, the bottom most detection scale is extended with a convolution layer, a max pooling layer and a convolution layer for classification.

4. Experiments

We performed various experiments to show how our method can handle the image registration challenges encountered in an in-vivo setting. For this we used data from our previous work with various visibility conditions. For training, we selected two sets of data, an ex-vivo setting and an in-vivo setting, including both nominal conditions for yellow and green amniotic fluid. For each setting a minimum of 25 and a maximum of 42 images were obtained for the same trajectory on the placenta. The number of images vary because of the differences in visibility. In total 745 images were used for various settings.

The training data was augmented by rotating the image in steps of 45 degrees and flipping it, such that 16 variations are obtained. For testing, all variations in visibility are used. Therefore, in nominal conditions 20% of the total set is used for testing and the rest is used for training. For all other visibility conditions all data is used for evaluation.

4.1. Experiment 1—Stable Region Detector

The stable region detector as proposed in Section 3.1 should detect the center of the vein. Therefore, we manually annotated the center and the radius of the veins and extracted the direction of the veins. According to the chosen scales and number of cells in the convolutional layers, the closest annotated point is selected as the ground truth and used to train the stable region detection network.

We evaluated the detection performance of these regions as well as their reproducibility for both the bounding boxes (BBox) and rotated boxes (RBox). We applied a confidence threshold of 0.95 and obtained on average 21.0, with a minimum of 11, regions per image in the in-vivo setting. With this high threshold the performance is also very high with 94.4% correctly detected regions. Lowering the threshold provides more regions albeit that the precision goes down very quickly. Below 0.7 only incorrect regions are detected. Therefore, we used this threshold of 0.95 in the rest of our experiments. The results of the BBox detections with more thresholds are shown in Table 1.

Table 1. Number of detections and precision.

Threshold	Ex-Vivo		In-Vivo	
	BBox	**RBox**	**BBox**	**RBox**
0.95	23.7 96.6%	25.8 97.1%	18.1 92.9%	21.0 94.4%
0.90	26.9 89.2%	29.2 91.0%	21.1 85.5%	24.8 86.6%
0.80	34.8 73.5%	35.9 80.1%	27.4 71.2%	30.8 73.5%
0.70	41.0 63.4%	43.9 66.7%	39.0 52.4%	40.5 59.1%
0.60	52.8 49.2%	58.8 55.3%	50.7 41.7%	51.8 50.5%

To determine the reproducibility of the detected regions, the transform between two successive images have been manually established. The ratio of the detections in two adjacent images that describe the same area are obtained by transforming the detections from one image to the other. The reproducible number of detections is on average 81.8% of the detected regions for the ex-vivo, and for the in-vivo settings 76.5% and 73.6%. For all visibility conditions an overview is presented

in Table A1. It also provides a comparison with the results of the key-point methods from our previous work.

4.2. Experiment 2—Stable Region Matching

To obtain matchable features we trained the neural network with Contrastive Loss on the matches. To evaluate the matching performance of our approach, the true matches from the previous experiment are used and compared to the number of matched regions. For the nominal ex-vivo setting 73.4% and for the nominal in-vivo settings 69.3% and 58.4% were correctly matched. For these settings all images had enough stable matches to obtain image registration. Furthermore, the mean pixel error was less than 2 pixels using LMeDS as the transform estimation method. Table A2 shows the matching performance for the other more challenging settings than nominal. For some visibility conditions an insufficient ratio of correct matches were found to use LMeDS, thus RANSAC was used instead.

4.3. Experiment 3—Qualitative Measure

To obtain a qualitative measure for the matching process as a whole for two adjacent images, the performance of the previous experiment is defined as *bad* if no transformation could be found either by having not enough detected regions or having not enough correctly matched pairs. A *good* performance is defined by more than 50% correct matches and a minimum of 6 correct matches. Which is based upon the requirement of LMeDs of having at least 50% correct matches and having more than 4 matches to handle location inaccuracy. By training an output with these labeled outcomes a measure of matchability could be obtained.

To obtain a qualitative measure of the visibility, a dataset was created containing also the *dark*, *light*, *close* and *far* visibility conditions. For the illumination and distance variation the nominal situation was defined as 0 and the two extremes of the variation as either -1 or 1 and trained with Euclidean loss. Table 2 shows the results for the qualitative measures as a ratio of giving a correct indication and an overall correct indication of successful image registration, where these measures are combined for the nominal setting.

Table 2. Qualitative Measure Precision.

Measure	Variation	Ex-Vivo	Yellow	Green
Distance	close	65%	60%	42%
	nominal	70%	68%	58%
	far	76%	72%	61%
Illumination	dark	88%	76%	40%
	nominal	90%	82%	60%
	light	93%	87%	83%
Matching		98%	95%	88%
Registration		100%	98%	91%

5. Discussion

In this paper we proposed an extension of an SSD network to detect regions in fetoscopic images with stable matchable features. With the same network architecture we also obtain a measure of matchability for the purpose of obtaining a sufficient set of matchable regions of consistent quality for proper image registration.

In Experiment 1 we showed that it is possible to detect stable regions on the placenta based on the medial axis of veins, under visibility conditions encountered in an in-vivo setting. Compared to key-point methods our approach only detects a limited number of regions, albeit that the number of reproducible regions is much higher and more consistent over all adjacent images in a trajectory.

The method to learn matchable feature was evaluated in Experiment 2. It showed that in better visibility conditions, a high percentage of correct matches could be obtained and that the number of correct matches is especially in darker settings reduced. Therefore, in the more complicated settings sometimes not enough matches could be found to obtain a transformation. However, in the nominal settings for 100% of the images sufficient matches could be found to obtain a transform.

These results show again that the visibility greatly complicates the in-vivo setting. First, for both the yellow and green-turbid liquid, the darker conditions have not enough contrast to provide the required detail to detect enough regions and extract matchable features. Next, the distance to the placenta also reduces the amount of regions that can be detected, resulting in not enough matches to either use LMeDS or obtain a transform estimation. Last, for the green-turbid settings, many images contain a large amount of sensor noise. These images provide a large number of key-points, however with our region detection method, almost no stable region could be detected.

The transform estimation precision is not as accurate as we expected. It seems that also our region detection method does not describe the same physical location uniquely enough. A more accurate transform estimation should be obtainable with dense optimization. This is anyway required for panorama reconstruction of large sequences without loops.

As stated it will still be very difficult to estimate a correct transform for all different visibility conditions. Therefore, in these cases it is important to be able to detect that the visibility condition is not suited for image registration. Experiment 3 evaluates the three qualitative measures defined and their combination for image registration. In most cases it is possible to detect whether the image is suitable for image registration. Furthermore, as the visibility condition is of great influence on the construction of the panorama image, this visibility should be communicated to the surgical team such that they can adjust the visibility at certain points on the panorama.

6. Conclusions

The aim of this work was to improve the panorama reconstruction process for in-vivo fetoscopic imaging. Our starting point was given by the four recommended points of improvement as described in our previous work. First, we improved the key-point detection and matching method, by extending the SSD method to detect stable regions and extract matchable features. Next, the panorama reconstruction process was improved, by detecting the complicating visibility conditions for the image registration and discarding improperly matched image pairs. Furthermore, a measure of the visibility condition was extracted such that it can be fed back to the surgical team. In this way, fetoscopic images of higher matchability might be obtained by a retry of the surgical team. These three points of improvement make a crucial step in the direction of in-vivo panorama reconstruction. Note that, the point of improving the equipment itself was not discussed in this work. A major improvement will come from an increased field of view while keeping the diameter of the fetoscope as small as possible.

Author Contributions: Floris Gaisser conceived the presented concept and developed the neural network. Floris Gaisser and Boris Lenseigne performed the experiments. Suzanne Peeters and Dick Oepkes provided and evaluated the experimental setup. Pieter Jonker and Dick Oepkes supervised the findings of the work. Floris Gaisser and Pieter Jonker wrote and all reviewed the manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

TTTS	Twin-To-Twin Transfusion Syndrome
LMeDS	Least Median of Squares
RANSAC	Random Sample Consensus
RCNN	Region-based Convolutional Neural Network
SSD	Single Shot (multibox) Detector
BBox	Bounding Box
RBox	Rotated Box
ex-vivo	Setting outside of and without mimicking the human body
in-vivo	Setting inside of or (realistically) mimicking the human body
setting	Environmental condition of the (experimental) setup
visibility condition	Changable situation influencing the visibility of the fetoscope
key-point	Interesting point as detected by methods like SIFT, SURF etc.

Appendix A. Visibility Conditions

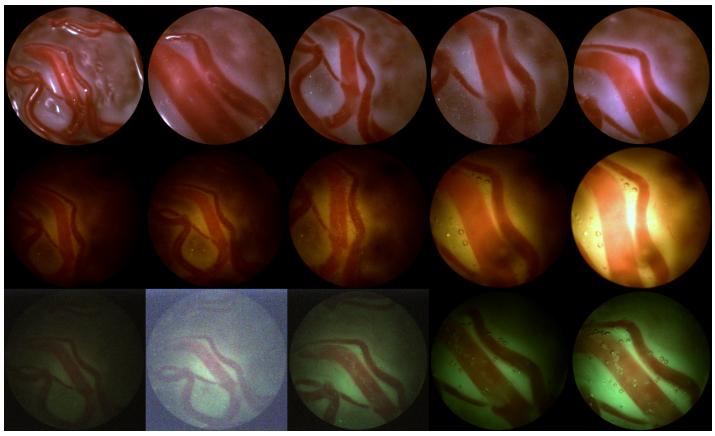

Figure A1. Variations in viewing conditions. Top row left to right: ex-vivo-far, ex-vivo-close, ex-vivo with water-far, ex-vivo with water-nominal, ex-vivo with water-close; Middle row: yellow liquid, bottom row: green turbid liquid; left to right: far-dark, far-nominal, nominal for both, close-nominal, close-bright.

Appendix B. Detailed Results

Please note that the results of the Key-point methods have been copied from Experiments 1 and 2 from previous work.

Table A1. Experiment 1: Key-points/regions detected.

Setting	Condition		Detected					Reproducible				
			SIFT	SURF	ORB	BBox	RBox	SIFT	SURF	ORB	BBox	RBox
ex-vivo	nominal		269	643	470	23.7	25.8	10.9%	23.3%	17.6%	80.1%	**81.8**%
Yellow	dark	close	14	31	21	8.6	9.1	13.1%	21.5%	27.7%	30.1%	32.5%
	dark	nominal	7	26	5	10.2	10.7	15.4%	31.4%	8.4%	33.0%	31.9%
	dark	far	22	50	15	11.2	12.8	15.4%	26.2%	15.8%	34.4%	37.0%
	nominal	close	42	132	57	12.9	13.2	23.3%	25.4%	34.8%	62.7%	62.8%
	nominal	nominal	24	110	21	20.4	23.4	16.5%	25.0%	21.1%	73.6%	**76.5**%
	nominal	far	74	174	73	20.7	23.8	17.6%	25.2%	23.8%	70.1%	71.2%
	light	close	97	350	108	12.0	11.9	22.4%	27.0%	35.3%	60.4%	60.1%
	light	nominal	110	525	129	16.1	17.3	19.5%	33.8%	23.3%	65.6%	66.7%
	light	far	38	183	32	18.7	18.8	13.2%	24.1%	15.7%	69.6%	70.0%

Table A1. *Cont.*

Setting	Condition		Detected					Reproducible				
			SIFT	SURF	ORB	BBox	RBox	SIFT	SURF	ORB	BBox	RBox
ex-vivo	nominal		269	643	470	23.7	25.8	10.9%	23.3%	17.6%	80.1%	**81.8**%
Green	dark	close	24	11	25	4.6	4.9	27.9%	25.3%	33.4%	5.1%	4.8%
	dark	nominal	11	13	10	4.5	4.6	27.2%	26.1%	15.6%	6.0%	6.1%
	dark	far	210	192	54	2.7	2.8	4.2%	7.3%	0.5%	1.3%	1.3%
	nominal	close	94	147	137	11.7	13.8	26.1%	31.4%	30.9%	57.5%	61.1%
	nominal	nominal	239	401	98	15.8	18.6	11.3%	24.4%	15.5%	68.6%	**73.6**%
	nominal	far	1000	1000	776	3.1	2.8	17.4%	23.0%	18.6%	2.5%	2.6%
	light	close	246	434	320	9.3	9.7	32.6%	36.7%	34.6%	55.9%	60.8%
	light	nominal	1000	1000	578	2.3	2.7	19.9%	18.4%	21.4%	2.6%	2.1%
	light	far	1000	1000	844	1.8	2.4	19.9%	23.2%	22.2%	2.4%	2.5%

Table A2. Experiment 2: Matches found and pixel error with LMeDS; * with RANSAC.

Setting	Condition		Correctly Matched		Sufficient Matches		Pixel Error	
			BBox	RBox	BBox	RBox	BBox	RBox
ex-vivo	nominal		71.3%	73.4%	100%	100%	2.1 ± 0.8 px	$\mathbf{1.9 \pm 0.7}$ **px**
Yellow	dark	close	24.7%	25.7%	0.0%	0.0%	−−	−−
	dark	nominal	25.8%	25.8%	0.0%	0.0%	−−	−−
	dark	far	31.7%	32.4%	0.0%	0.0%	−−	−−
	nominal	close	45.5%	45.8%	39.8%	40.9%	3.1 ± 1.4 px *	3.1 ± 1.3 px *
	nominal	nominal	65.2%	69.3%	100%	100%	2.0 ± 0.9 px	$\mathbf{1.9 \pm 0.8}$ **px**
	nominal	far	63.7%	67.1%	100%	100%	2.1 ± 0.8 px	1.9 ± 0.6 px
	light	close	42.6%	42.8%	32.6%	34.9%	3.4 ± 1.3 px *	3.2 ± 1.3 px *
	light	nominal	55.1%	55.9%	96.6%	100%	2.4 ± 1.1 px	1.9 ± 0.7 px
	light	far	58.3%	60.4%	100%	100%	2.1 ± 0.8 px	1.9 ± 0.7 px
Green	dark	close	−−	−−	−−	−−	−−	−−
	dark	nominal	−−	−−	−−	−−	−−	−−
	dark	far	−−	−−	−−	−−	−−	−−
	nominal	close	41.7%	49.3%	2.4%	47.6%	−−	3.2 ± 1.2 px *
	nominal	nominal	55.7%	58.4%	100%	100%	2.5 ± 0.9 px	$\mathbf{2.1 \pm 0.8}$ **px**
	nominal	far	−−	−−	−−	−−	−−	−−
	light	close	35.4%	35.5%	0.0%	2.4%	−−	−−
	light	nominal	−−	−−	−−	−−	−−	−−
	light	far	−−	−−	−−	−−	−−	−−

References

1. Lewi, L.; Deprest, J.; Hecher, K. The vascular anastomoses in monochorionic twin pregnancies and their clinical consequences. *Am. J. Obstet. Gynecol.* **2013**, *208*, 19–30.
2. Peeters, S. Training and Teaching Fetoscopic Laser Therapy: Assessment of a High Fidelity Simulator Based Curriculum. Ph.D. Thesis, Leiden University Medical Center, Leiden, The Netherlands, 2015.
3. Seshamani, S.; Lau, W.; Hager, G. Real-time endoscopic mosaicking. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2006*; Springer: Berlin, Germany, 2006; pp. 355–363.
4. Soper, T.D.; Porter, M.P.; Seibel, E.J. Surface mosaics of the bladder reconstructed from endoscopic video for automated surveillance. *IEEE Trans. Biomed. Eng.* **2012**, *59*, 1670–1680.
5. Carroll, R.E.; Seitz, S.M. Rectified surface mosaics. *Int. J. Comput. Vis.* **2009**, *85*, 307–315.
6. Tella-Amo, M.; Daga, P.; Chadebecq, F.; Thompson, S.; Shakir, D.I.; Dwyer, G.; Wimalasundera, R.; Deprest, J.; Stoyanov, D.; Vercauteren, T.; 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, Seattle, WA, USA, 27–30 June 2016; pp. 84–92.
7. Gaisser, F.; Jonker, P.P.; Chiba, T. Image Registration for Placenta Reconstruction. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 27–30 June 2016; pp. 33–40.

8. Reeff, M.; Gerhard, F.; Cattin, P.C.; Székely, G. Mosaicing of endoscopic placenta images. In *Informatik für Menschen*; Hartung-Gorre Verlag: Konstanz, Germany, 2006.
9. Liao, H.; Tsuzuki, M.; Kobayashi, E.; Dohi, T.; Chiba, T.; Mochizuki, T.; Sakuma, I. 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*; Springer: Berlin, Germany, 2008; pp. 329–338.
10. Rousseeuw, P.J. Least Median of Squares Regression. *J. Am. Stat. Assoc.* **1984**, *79*, 871–880.
11. Fischler, M.A.; Bolles, R.C. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM* **1981**, *24*, 381–395.
12. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444.
13. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* **2014**, arXiv:1409.1556. Available online: https://arxiv.org/abs/1409.1556 (accessed on 12 January 2018).
14. Verdie, Y.; Yi, K.; Fua, P.; Lepetit, V. TILDE: A temporally invariant learned detector. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 5279–5288.
15. Simo-Serra, E.; Trulls, E.; Ferraz, L.; Kokkinos, I.; Fua, P.; Moreno-Noguer, F. Discriminative learning of deep convolutional feature point descriptors. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 118–126.
16. Yi, K.M.; Trulls, E.; Lepetit, V.; Fua, P. Lift: Learned invariant feature transform. In *European Conference on Computer Vision*; Springer: Berlin, Germany, 2016; pp. 467–483.
17. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Washington, DC, USA, 23–28 June 2014; pp. 580–587.
18. Girshick, R. Fast R-CNN. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
19. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*; IEEE: Piscataway Township, NJ, USA, 2015; pp. 91–99.
20. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. SSD: Single shot multibox detector. In *European Conference on Computer Vision*; Springer: Berlin, Germany, 2016; pp. 21–37.