Image Registration for Placenta Reconstruction

Floris Gaisser
f.gaisser@tudelft.nl

Pieter P. Jonker
p.p.jonker@tudelft.nl

Delft University of Technology
Department of BioMechanical Engineering

Toshio Chiba
chiba-t@sea.plala.or.jp

Nihon University
University Research Center

Abstract

In this paper we introduce a method to handle the challenges posed by image registration for placenta reconstruction from fetoscopic video as used in the treatment of Twin-to-Twin Transfusion Syndrome (TTTS). Panorama reconstruction of the placenta greatly supports the surgeon in obtaining a complete view of the placenta to localize vascular anastomoses. The found shunts can subsequently be blocked by coagulation in the correct order.

By using similarity learning in training a Convolutional Neural Network we created a novel feature extraction method, allowing robust matching of keypoints for image registration and therefore taking the most critical step in placenta reconstruction from fetoscopic video.

The fetoscopic video we used for our experiments was acquired from a training simulator for TTTS surgery. We compared our method with state-of-the-art methods. The matching performance of our method is up to three times better while the mean projection error is reduced with 64% *for the registered images. Our image registration method provides the ground work for a complete panorama reconstruction of the placenta.*

1. Introduction

The Twin-to-Twin Transfusion Syndrome (TTTS) is a condition that occurs in about 10% of the pregnancies involving monochorionic twin (twins with a shared placenta). In TTTS an unbalanced exchange of blood caused by vascular anastomoses (shunts) in the placenta causes fatal complications for both twins [11]. Fetoscopic surgery is a common technique used to separate the fetal circulations by coagulating the connecting vessels with a laser beam. This technique has shown large improvements in the survival rate over other treatments [5]; even though far from optimal, it is currently a widely applied procedure [14].

One of the limitations that complicates the surgery is the very limited view of the fetoscope and the lack of a complete overview as the fetus is generally occluding large parts of the placenta. Limitations on the endoscope diameter limit the possibilities to improve this on the level of the surgical instruments. A complete view of the placenta would greatly support the surgeon in localizing all vascular anastomoses and guiding the surgeon to coagulate these shunts in the correct order [14].

To obtain such a view, this paper presents a novel method for image registration as this is the most critical step in the construction of a panorama from fetoscopic video. Our approach uses similarity learning at training a Convolutional Neural Network (CNN), to create a feature extractor that is suited to the images from the fetoscope and invariant to the transformations encountered by its movements. This feature extractor allows both robust matching of keypoints and transform estimation.

2. Related Work

Reconstructing large view panoramas of the internal anatomical structures has been a large field of research and found many applications, such as retina [18], bladder [19] and oesophagus [4] reconstructions, as well as in ex-uterin endoscopic mosaicking [17, 12].

First, [17] shows the reconstruction of a small part of an ex-vivo placenta, though the results show that the image registration has a low accuracy and the reconstruction without post-processing contains many artefacts. Furthermore, the images are captured by moving the camera sideways in a structured circular pattern. First of all this cannot be reproduced in an in-vivo setting, but also the transforms between images now only consists of translations.

Second, in [12] they project endoscope images of a color injected placenta on a 3D ultrasound model, which shows accurate results in image registration. However, such a setting with an ex-vivo color injected placenta is not compatible with our goal of in-vivo surgery.

Third, there have been promising results in other applications such as bladder reconstruction [19]. There an ex-vivo dye injected bladder is reconstructed from a flexible endoscope with image registration, bundle adjustment and spherical projection. However, also here this method is not

suited for our setting, as no prior structure is available. Furthermore, the encountered transformations here and in [18] are also mostly translations which can be robustly estimated with existing methods.

Last, in oesophagus reconstruction [4] an accurate reconstruction is presented, however here pipe projection is used, which is also not applicable to our setting. Furthermore, spatial consistency is not required for this type of reconstruction.

Although all above methods are not directly applicable in our aimed setting, some successes have been shown.

2.1. Image Registration

The previously discussed applications all use image registration methods which try to find the transformation between two images [8]. They try to find corresponding pairs of interesting points in both images by feature matching, whereafter a transform is estimated based on the found matches [22].

To find matching pairs, first interesting keypoints are chosen using methods such as the maximum Difference of Gaussians [13] as used in SIFT. Next, to find the corresponding point in the other image, the selected keypoints are described using a feature extraction method, such that the features are similar regardless of the appearance changes due to the transforms between the images. Obtaining such features has been the source of many invariant methods such as the Scale-Invariant Feature Transform (SIFT) [13] or Binary Robust Independent Elementary Feature (BRIEF) [3].

Though feature extraction methods are designed to be invariant to transformations, there are still challenges in obtaining appropriate matches. To handle incorrect matches, transform estimation methods try to find a best fitting estimation by iteratively fitting on random subsets of the matches and selecting the best fitting subset. RANSAC [6] is robust to mismatches but finds a sub-optimal estimation, where LMedS [16] finds a more accurate estimation but requires at least 50% correct matches.

2.2. Problem statement

Our initial as well as other research [17, 19] showed that the state-of-the-art methods have promising results but lack application in a realistic setting, i.e. it cannot be applied in real surgery. Hence, our research focusses on using fetoscopic videos from a more realistic setting which introduces challenges not encountered before.

First of all, there is the loss of contrast of the blood vessels due to the inability to use dye injected placentas. Then, most of the time complex perspective transforms are encountered as the endoscope has a fixed point entering the uterus and the view is mostly changed by rotating about this entry point. Finally, since reconstruction of the placenta has to be done near real-time, long post-processing is not possible and therefore the transform estimation has to be fairly accurate and also consistent.

Our initial research showed that on our fetoscopic images, state-of-the-art keypoint methods fail to extract robust keypoints and features, partly because these methods are designed for natural images and require unique and distinctive structures. But in our case blood vessels on the placenta are very similar and have a very limited structure.

2.3. Convolutional Neural Networks

In the fields of Machine Learning and Computer Vision, deep-learning neural networks have found a wide range of applications due to their ability to learn specific concise representations from the raw image data [10, 20]. They outperform many state-of-the-art methods as well as the previously described keypoint description methods. Furthermore, inspired by the neural sciences on how humans learn, a Convolutional Neural Network (CNN) can be trained to extract invariant features by using similarity learning [7, 21]. Consequently, these characteristics motivate us to use CNNs to cope with the challenges encountered in fetoscopic image registration.

3. Method

In contrast to keypoint feature extraction methods, convolutional neural networks have to be trained to learn a mapping between the input image data and a feature vector. Our proposed method uses a two staged approach; first a network is trained to extract features that are robust to small perspective transforms. Second, training an extension of this first network is performed to fine-tune the feature extraction, in order to obtain features that can be matched robustly.

To train any neural network, a loss function is used to acquire the feedback for updating the internal state of the network. Our method is described in detail in section 3.1. The network for image registration is described in section 3.2 detailing the feature extraction and the matching and registration parts of the network. As the network is trained using a training set, the creation of the training set is described in section 3.3. Finally, the remaining sections describe the experimental setup (3.4), the results (4) and a discussion of the results (5).

3.1. Learning Method

CNNs learn a mapping between the input and required output by updating internal weights based on feedback given to the network. This feedback, also defined as the error or the loss, is obtained by defining a function which generally takes the current and the desired output of the network as inputs. This function tries to minimize the error between output of the network and desired output, thus using only feedback on similarity.

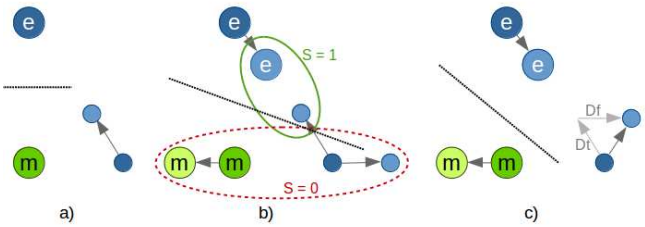

Figure 1. (e) elephant label / features, (m) mouse label / features a) Label based learning b) contrastive loss c) matching learning

A different approach is to also define feedback on dissimilar inputs. This is achieved with the contrastive loss function [7]. Which defines feedback to decrease the difference between similar pairs and to increase the difference between dissimilar pairs, which results in a more easily separable and more evenly distributed feature space.

To describe the difference in feedback, consider a network trained to classify images containing either a mouse or an elephant. Suppose during training a sample of an elephant is incorrectly described as a mouse. Normally feedback is provided to decrease the difference between the class label from the network and the label from the training sample. This results in making the output more similar to the elephant label, as shown in figure 1a where the feature after learning is still closer to the incorrect mouse label.

For contrastive loss training, a siamese network [1] using two images is used to train the network. Generally, this method is utilized to train a network for feature extraction making the output a feature vector. In the case where a sample of an elephant and a mouse is used, the difference between their outputs is increased up to a defined margin, as shown in figure 1b with the red dashed ellipse. However, in the case two samples of the same label are used, the difference between the two outputs is minimized as shown with the green solid ellipse. Hence improving the feature extraction towards their correct label, as well as making the two features more dissimilar and more easily separable.

Our goal is to train a CNN to extract invariant and robust features to describe key areas. To realise invariance to perspective transforms, the error between different transformations of the same patch has to be minimized, while to extract features that are separable, the error between different patches has to be maximized. This can be achieved with the contrastive loss function as is defined in (1). Where X_i is the output of the network as feature vectors, m the margin, generally defined as 1, s the similarity of the pair with 1 as similar and 0 as dissimilar. For more details we refer to the original work on contrastive loss [7].

$$L = s\frac{1}{2}(D_w)^2 + (1-s)\frac{1}{2}(max(0, m - D_w))^2 \quad (1)$$

$$D_w = \|X_1 - X_2\|_2$$

In the process of image registration, extracted features are matched on their Euclidean norm similarity. To train a network to extract features that can be matched robustly, the contrastive loss function is extended. The ground truth from the training samples is used to select true matches. Next, feedback is defined such that the error between incorrectly matched features is increased and between correctly or supposedly matched features is decreased. This is described by (2), where $f = 1$ when the feature matching obtained a false match and $f = 0$ when the feature matching was correct. D_f and D_t are respectively the differences between X_1 and the feature vector obtained by feature matching X_f or X_t obtained by the true transform.

$$L = \frac{1}{2}((1-f)D_f + fD_t)^2 + f\frac{1}{2}(max(0, m - D_f))^2 \quad (2)$$

$$D_f = \|X_1 - X_f\|_2$$
$$D_t = \|X_1 - X_t\|_2$$

Function (2) is inspired by the contrastive loss function, in minimizing the difference between correct matches and increasing the difference between incorrect matches. But it differs by introducing two reference features to match with; the true match X_t and the feature based match X_f. In the case where the feature matching was correct ($f = 0$), these two references are the same, and (2) can be considered similar to the case where $s = 1$ in (1) as the second term is cancelled out. However, in the case where the feature matching obtained an incorrect match ($f = 1$), additional feedback is given based on the incorrect match. This has as effect that not only the correct features are made more similar and the incorrectly matched features more dissimilar, but also that the specific aspects that form the difference between the correct feature and the incorrect feature are improved.

To describe this effect, consider the previous example of training a network describing images of a mouse and an elephant. Imagine the feature vector describing some aspects of the animals including colour and size. Suppose during training an image of an elephant was mistakenly matched with a feature of a mouse. The feedback will increase the difference between these two features, in both colour and size. Furthermore, feedback is given to reduce the difference between the correct feature and the extracted feature. As the size of an elephant is large, the aspect of size is increased even more. But as both animals are grey, the aspect of colour is reduced. Even so, the importance of the colour aspect is reduced over time up to the point that the network will not use colour any more to describe the animals. This is shown in figure 1c with the combination of the difference between the incorrect feature and extracted feature D_f as well as the difference between the correct feature and the correct feature D_t. Resulting in a much better separable feature space as indicated with the black dotted line.

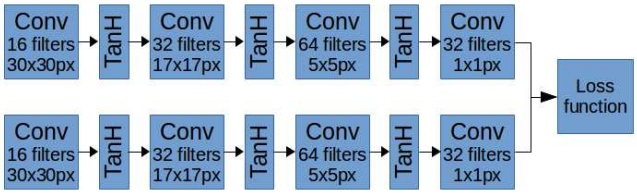

Figure 2. CNN architecture.

It can be argued that the triplet learning from [21] is very similar to our proposed method. However, there is one key difference in the way how a dissimilar pair is chosen. In [21] this is a fixed pair chosen at the moment the training set is created, whereas our method dynamically obtains a dissimilar pair based on the output of the network. Therefore it is adaptive to what is learned in the network, creating a much better separable feature space. Furthermore no dissimilar pairs have to be selected when creating a training set, reducing the training set size as well as training time significantly.

3.2. Network Architecture

As stated before, the network is trained in two stages; feature extraction training and robust matching training. Both stages use a siamese network architecture, where two parallel networks with the same architecture share their internal weights to process two simultaneous inputs [1].

For feature extraction, a network is designed such, that an input image patch of 50×50 pixels is reduced to a feature vector of size 32, by choosing the right number and filter sizes for the convolution layers as shown in figure 2.

For training robust matchable features, the same network is used, but instead of a single image patch, 961 patches of 50×50 pixels are extracted in a 31×31 grid from a 500×500 image. Furthermore the contrastive loss layer is replaced with the matching loss layer as described in the previous section.

For evaluation with image registration, the matching loss layer is replaced with a matching and rigid transform estimation layer. This layer outputs the estimated rigid transform found by RANSAC or LMEDS [6, 16] and the mean projection pixel error between the true transform and the estimated transform.

Algorithm 1 Training data

Step 1: Create image patches.
Step 2: Discard similar patches
Step 3: Select only interesting patches
Step 4: Create transformed patches
Step 5: Similarity pairing

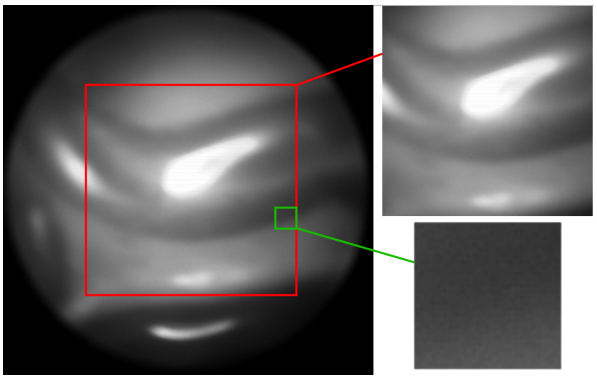

Figure 3. Image from fetoscope and crops for learning

3.3. Training data

To train any CNN, a dataset has to be created that is as small as possible to reduce the training time. As well as a complete and an evenly distributed representation of the variations to be encountered, in order to achieve robustness and avoid over-fitting. In algorithm 1 the steps for creating these training sets are shown and detailed below.

First, a subset of images from the fetoscopic videos are selected to decrease the amount of training data. As the motion within one second is expected to be small, only 5 images each second are selected. Next, for the first training stage, patches of 50×50 pixels (figure 3 right bottom) are extracted and for the second stage patches of 500×500 pixels (right top) are extracted at an interval of 50 pixels from the valid area of 550×550 pixels of the source images.

Steps 2 and 3 are to improve the quality of the extracted patches used in the dataset. First, the absolute pixel difference between all patches is obtained. Patches that are too similar are discarded, such that reoccurring variations are not presented multiple times. As a result, the dataset contains an evenly distributed representation of the variations. To further improve the information density of the dataset, all patches with below average gradient energy are discarded. This results in a set of patches that are above average descriptive and makes sure that non-descriptive patches are excluded.

In order to have invariance to the expected transformations, every patch is rigidly transformed. For the training sets, fixed step sizes are chosen for every component of the perspective transform, related to the observed transforms occurring between two successive frames. Similarly, for the evaluation sets, random transforms are chosen.

For similarity training, pairs are created in the final step where every patch is paired with their variations. Furthermore for the first training stage also dissimilar pairs have to be selected. Therefore every patch is paired with 25 of their most similar patches based on the absolute pixel difference obtained in step 2.

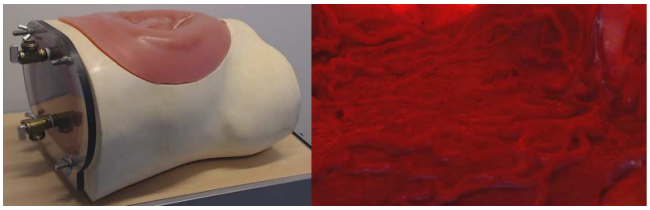

Figure 4. Left: Simulator, Right: inside of simulator with placenta

3.4. Experimental Setup

To evaluate the introduced image registration method, fetoscopic videos were utilized from a TTTS surgery simulator used to train surgeons as shown in figure 4 [15].

It has to be noted that the artificial model of the placenta as shown in figure 5, is a close as possible representation of a real placenta. This is unlike the much easier dye injected placentas that are used in the current state-of-the-art. Furthermore, the positioning of the placenta and use of the fetoscope is similar to that of in-vivo surgery (figure 4).

The image registration method has been implemented on a Dell precision M4700 with the Caffe [9] and OpenCV libraries. The videos have been acquired with a medical camera capturing a circular image of 880×880 pixels representing an area of about 8×8 mm as shown in figure 3.

4. Results

For performance evaluation of image registration in a realistic setting, a video taken from the simulator operated by an expert is processed. Three sections of the video have been chosen with similar length of about 75 seconds, representing different areas of the placenta. Training is performed on one of the videos and compared with the other two. Patches are extracted for both the first and second stage of training and evaluation as described in section 3.3. By changing the training set, three combinations of training and evaluation could be obtained.

4.1. Experiment 1

First the invariance of the novel feature extraction method is evaluated in respect to the different transformations and compared to the state-of-the-art keypoint descriptors. In table 1, the average performance is shown together

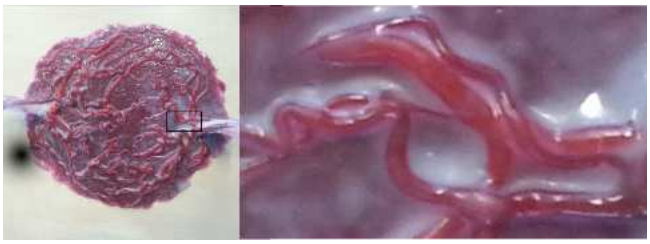

Figure 5. Artificial placenta

Table 1. Correctly matched points

Method	SIFT	BRIEF	CNN1	CNN2
Translation	28.2%	29.5%	67.5%	**81.4%**
	±24.0%	±10.1%	±15.6%	**±13.6%**
Rotation	22.1%	31.5%	53.4%	**74.5%**
	±19.0%	±7.8%	±13.1%	**±12.6%**
Scale	21.1%	36.1%	57.8%	**72.9%**
	±16.9%	±7.9%	±11.6%	**±12.2%**
Perspective	13.7%	27.4%	51.4%	**68.4%**
	±9.8%	±7.2%	±7.1%	**±12.2%**
All	13.8%	26.2%	50.9%	**62.8%**
	±4.7%	±6.7%	±3.9%	**±6.1%**

with the standard deviation of the correctly matched points out of the total keypoints. CNN1 represents the performance trained only with the first stage, while CNN2 was trained with the novel matching learning method.

All methods use a fixed grid of 31×31 points with a spacing of 10 pixels, therefore always having 961 keypoints for feature extraction. This was also chosen for SIFT and BRIEF to guarantee that keypoints were available that represented the same area in both images. For both SIFT and BRIEF, a match was only accepted if the distance ration to the second best match was below a threshold as shown in [2]. This threshold was adjusted such that only the best matches, but also enough matches could be retained for the next experiment.

4.2. Experiment 2

For performance evaluation of image registration in a realistic setting, comparable to in-vivo surgery, table 2 shows the image registration error as the mean pixel error of the estimated transform together with the standard deviation.

For state-of-the-art keypoint description methods, RANSAC is used for transform estimation, whereas for the proposed methods also LMedS is used, as more than 50% of the matches are correct matches.

It should be noted that even by adjusting the threshold, for both SIFT and BRIEF, in 10-25% of the images the matching ratio was so low that less than the required 4 matches were found. Furthermore, for about 15-25%, no reasonable transform estimation could be found. These have all been excluded from this comparison, as they influenced the average pixel error drastically.

4.3. Experiment 3

Using 26 sequential registered images from the previous experiment, a partial reconstruction of the placenta, as shown in figure 6, has been made of the same area shown in figure 5. In this reconstruction, no post-processing or blending methods were used, but still giving promising results.

Table 2. Mean pixel error of estimated transform. [1)] RANSAC [2)] LMeDS

Method	SIFT	BRIEF	CNN1 [1)]	CNN1 [2)]	CNN2 [1)]	CNN2 [2)]
Translation	4.0 ±1.7 px	3.5 ±1.7 px	3.2 ±3.4 px	2.6 ±1.8 px	2.6 ±1.5 px	**2.4 ±1.4 px**
Rotation	7.1 ±1.9 px	8.0 ±2.0 px	6.6 ±5.0 px	4.0 ±2.6 px	4.3 ±2.8 px	**2.4 ±1.7 px**
Scale	7.1 ±1.8 px	8.6 ±1.4 px	5.3 ±3.7 px	3.6 ±2.0 px	4.6 ±3.1 px	**2.6 ±1.6 px**
Perspective	9.9 ±3.1 px	9.8 ±2.8 px	6.6 ±3.9 px	4.2 ±2.6 px	5.7 ±3.2 px	**2.9 ±1.6 px**
All	8.3 ±3.0 px	8.5 ±2.7 px	7.5 ±4.0 px	5.2 ±2.9 px	6.6 ±3.1 px	**3.0 ±1.6 px**

5. Discussion

In this paper an image registration method is introduced to handle the challenges posed by fetoscopic videos. The main challenge in image registration is to obtain invariant features that can be matched robustly. With the experiments it was shown that feature extraction with a CNN trained in a novel way, allows for more robust features and improves image registration of fetoscopic images.

The first experiment shows that depending on the applied rigid transformation, for the novel approach of using learned feature extraction, up to 67.5% of the features can be matched. The key behind this, is that the network learns to extract the essential components to describe an area, such that it is still invariant to the applied transforms.

The remarkable low matching performance of state-of-the-art methods can be explained by the ratio between the robustness to variations and the difference between different keypoints. For both SIFT and BRIEF, as they are designed to be invariant to these type of rigid transformations, the difference between extracted features of similar keypoints is small. Thus, for robust keypoint matching, it requires a very different type of keypoint, which is also the reason why it is advised to only accept matches by a distance-to-second-best ratio. However, as having different type of keypoints is not feasible with fetoscopic images, since blood vessels look very similar, the result is a low matching performance.

This also explains the two causes why many of the matching samples for SIFT and BRIEF had to be excluded from the results. This had two causes. First, the distance-to-the-second-best-ratio threshold rejects the majority of matches, resulting in less than 4 matches. Second, the features are too similar and are matched incorrectly.

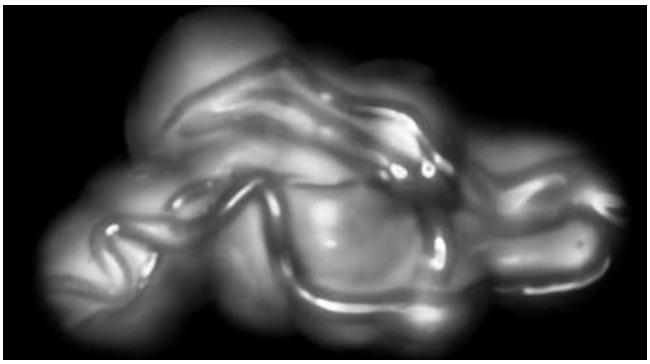

Figure 6. Reconstruction of placenta

In contrast to state-of-the-art keypoint descriptors, our novel matching learning method increases the difference between different areas on top of the invariant feature extraction. This is shown in the improvement in matching performance between CNN1 50.9% and CNN2 62.8% for all transforms.

In [17] they showed a matching performance of 68% for SIFT matching. This is quite different from the results presented in this paper. But, this difference can be explained by three aspects. First, the field of view of their endoscope is larger, showing much more structure. Second, they use a dye injected placenta, which results in much more contrast allowing for better features to be extracted. Third, the motion they used during recording consists mostly of translation. SIFT obtains a 2 times better matching performance in experiment 1 for translation (28.2%) compared to the realistic transforms encountered during surgery which it only matches 13.8%.

The results of experiment 2 show that having more correct matches makes for more robust and precise transform estimation. This is reasonable because of the well known correlation between the amount of matches and the transformation error. It should also be noted that LMedS will give an optimal estimation, where RANSAC will give the best estimation of its iterations. This can be seen from the results of CNN1 with LMedS and CNN2 with RANSAC for all transforms. The latter has more correct matches, 62.8% compared to 50.9%, but also a higher estimation error of 6.6 compared to 5.2 pixels. Furthermore, looking at the individual matching results, it can be seen that RANSAC will sometimes give an estimation that is quite far off.

Another aspect that is often not considered is the consistency of the image registration process. With conventional keypoint matching methods some of the images could not be registered. The same problem has been reported in [17]. With our proposed method, 100% of the test images could be matched, as the features and matches obtained were very robust to the variations in the image data and the perspective transform between two successive images. Therefore, continuous and complete panorama reconstruction should be obtainable with this novel method.

In experiment 3 an attempt is made towards reconstructing large view panoramas, using images from a fourth sequence. Unfortunately, motion blur and lack of structure in small areas, limited the length of the sequence and therefore the area that could be reconstructed. However, the consistency of the obtained transform estimation shows that large view panoramas can be reconstructed. Furthermore shows that the quality of the videos is important as well.

One aspect of keypoint based image registration that is not covered in this paper is the detection of these keypoints. In this work, a grid of 31×31 is used as keypoints, where generally these are detected, such as in the detection part of SIFT. In future work, this aspect will also be included, but the exclusion of this aspect can be explained.

First, as stated before, it cannot be guaranteed that the detection will obtain keypoints that are matchable between the two images. In a grid of keypoints, this can be guaranteed with an increased distance, where the maximum possible distance between matchable keypoints, excluding the transformation, is half of the interval between the points on the grid.

Second, a placenta, consisting of a network of blood vessels, has very limited unique features. Moreover, the edge between a blood vessel and the underlying tissue of the placenta is very similar along the whole edge. As a result, a keypoint is generally arbitrarily detected along this edge and consistent keypoint detection cannot be guaranteed. For future work, a keypoint or an interesting area should be selected on the structure of this edge and not the gradient around a point on this edge.

6. Conclusion

In this paper, a novel method is described for the first and most crucial step in panorama reconstruction. This method can extract robust matchable features using a Convolution Neural Network, which is trained with a novel matching similarity learning method.

This approach largely improves the number of correctly matched features over the state-of-the-art methods. The feature matching, which is almost three times better, gives a 64% increase in transformation estimation accuracy. Furthermore, consistent registration can be achieved, because for every image, a reasonable transform estimation could be obtained, which is of great importance for the reconstruction of large view panoramas.

These improvements are achieved while fetoscopic images from a more challenging and realistic setting are used, in contrast to commonly used ex-vivo and dye injected settings. Furthermore, a partial reconstruction could be obtained, showing a 4 by 2 times larger view of the placenta while containing only minor visual artefacts.

Based on the obtained results, some recommendations can be made for future research. First, the detection of interesting points based on learned CNN for keypoint detection should be evaluated, as this is one of the main aspects that was not covered in this paper. Furthermore, on-the-fly bundle adjustment of the image registration using multiple locally overlapping images should be used to improve the reconstruction accuracy and its spatial consistency for a complete reconstruction of the placenta.

Furthermore recommendations towards implementation in surgery can be made. The current image data is captured from a single artificial model. The transferability of the trained network between the artificial model and an in-vivo setting as well between patients should be evaluated.

References

[1] J. Bromley, J. W. Bentz, L. Bottou, I. Guyon, Y. LeCun, C. Moore, E. Säckinger, and R. Shah. Signature verification using a siamese time delay neural network. *International Journal of Pattern Recognition and Artificial Intelligence*, 7(04):669–688, 1993. 3, 4

[2] M. Brown and D. G. Lowe. Automatic panoramic image stitching using invariant features. *International journal of computer vision*, 74(1):59–73, 2007. 5

[3] M. Calonder, V. Lepetit, C. Strecha, and P. Fua. Brief: Binary robust independent elementary features. *Computer Vision–ECCV 2010*, pages 778–792, 2010. 2

[4] R. E. Carroll and S. M. Seitz. Rectified surface mosaics. *International journal of computer vision*, 85(3):307–315, 2009. 1, 2

[5] R. H. Chmait, E. V. Kontopoulos, L. M. Korst, A. Llanes, I. Petisco, and R. A. Quintero. Stage-based outcomes of 682 consecutive cases of twin–twin transfusion syndrome treated with laser surgery: the usfetus experience. *American journal of obstetrics and gynecology*, 204(5):393–e1, 2011. 1

[6] M. A. Fischler and R. C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981. 2, 4

[7] R. Hadsell, S. Chopra, and Y. LeCun. Dimensionality reduction by learning an invariant mapping. In *Computer vision and pattern recognition, 2006 IEEE computer society conference on*, volume 2, pages 1735–1742. IEEE, 2006. 2, 3

[8] R. Hartley and A. Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003. 2

[9] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the ACM International Conference on Multimedia*, MM '14, pages 675–678, New York, NY, USA, 2014. ACM. 5

[10] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 2

[11] L. Lewi, J. Deprest, and K. Hecher. The vascular anastomoses in monochorionic twin pregnancies and their clinical consequences. *American journal of obstetrics and gynecology*, 208(1):19–30, 2013. 1

[12] H. Liao, M. Tsuzuki, E. Kobayashi, T. Dohi, T. Chiba, T. Mochizuki, and I. Sakuma. 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*, pages 329–338. Springer, 2008. 1

[13] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004. 2

[14] S. Peeters. *Training and teaching fetoscopic laser therapy: assessment of a high fidelity simulator based curriculum*. PhD thesis, Leiden University Medical Center, 1 2015. 1

[15] S. Peeters et al. Simulator training in fetoscopic laser surgery for twin–twin transfusion syndrome: a pilot randomized controlled trial. *Ultrasound in Obstetrics & Gynecology*, 46(3):319–326, 2015. 5

[16] Peter J. Rousseeuw. Least median of squares regression. *Journal of the American Statistical Association*, 79(388):871, , 880, December 1984. 2, 4

[17] M. Reeff, F. Gerhard, P. C. Cattin, and G. Székely. Mosaicing of endoscopic placenta images. In *Informatik fr Menschen*, volume 1. Hartung-Gorre Verlag, 2006. 1, 2, 6

[18] S. Seshamani, W. Lau, and G. Hager. Real-time endoscopic mosaicking. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2006*, pages 355–363. Springer, 2006. 1, 2

[19] T. D. Soper, M. P. Porter, and E. J. Seibel. Surface mosaics of the bladder reconstructed from endoscopic video for automated surveillance. *Biomedical Engineering, IEEE Transactions on*, 59(6):1670–1680, 2012. 1, 2

[20] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. *arXiv preprint arXiv:1409.4842*, 2014. 2

[21] P. Wohlhart and V. Lepetit. Learning descriptors for object recognition and 3d pose estimation. *arXiv preprint arXiv:1502.05908*, 2015. 2, 4

[22] B. Zitova and J. Flusser. Image registration methods: a survey. *Image and vision computing*, 21(11):977–1000, 2003. 2