Deep Learning for Sensorless 3D Freehand Ultrasound Imaging

Raphael Prevost[1], Mehrdad Salehi[1,2], Julian Sprung[3], Robert Bauer[3], and Wolfgang Wein[1]

[1] ImFusion GmbH, Munich, Germany
[2] Computer Aided Medical Procedures (CAMP), TU Munich, Germany
[3] piur Imaging GmbH, Vienna, Austria

Abstract. 3D freehand ultrasound imaging is a very promising imaging modality but its acquisition is often neither portable nor practical because of the required external tracking hardware. Building a sensorless solution that is fully based on image analysis would thus have many potential applications. However, previously proposed approaches rely on physical models whose assumptions only hold on synthetic or phantom datasets, failing to translate to actual clinical acquisitions with sufficient accuracy. In this paper, we investigate the alternative approach of using statistical learning to circumvent this problem. To that end, we are leveraging the unique modeling capabilities of convolutional neural networks in order to build an end-to-end system where we directly predict the ultrasound probe motion from the images themselves. Based on thorough experiments using both phantom acquisitions and a set of 100 in-vivo long ultrasound sweeps for vein mapping, we show that our novel approach significantly outperforms the standard method and has direct clinical applicability, with an average drift error of merely 7% over the whole length of each ultrasound clip.

1 Introduction

Ultrasound imaging (US) is one of the main medical modalities for both diagnostic and interventional applications thanks to its unique properties - affordability, availability, safety and real-time capabilities. For a long time though, its inability to acquire 3D images has reduced its range of clinical applications. The workaround was to acquire a series of 2D images by sweeping over the region of interest and combining them into a single volume afterwards. This solution requires the knowledge of the relative position from one image to the next. External sensor-based solutions (typically optical or electromagnetic) are only able to provide a good estimate of the probe position at the expense of practicality and price, while motorized or 2D array transducers have a limited field-of-view and are also quite expensive.

Thus, a significant amount of research has been dedicated at solving this problem without additional hardware by estimating the relative position of two images with pure image processing algorithms. While the in-plane motion can be

recovered quite reliably with algorithms like optical flow [1], the biggest challenge is to estimate the out-of-plane motion (often called *elevational displacement*). The reference approach exploits the very particular speckle noise patterns that are visible in ultrasound images, and is thus called *speckle decorrelation* [2, 3]. It is based on the fact that the US intensities undergo a point-spread function not only in the image plane but also in the perpendicular direction. This means that the speckle patterns of two successive frames have a strong correlation: the higher the correlation, the lower the elevational distance. Unfortunately, this relationship is far from trivial and a lot of papers have proposed various models based on the physical and statistical properties of the image acquisitions [2, 4]. While those methods produce fairly accurate estimates on synthetic data, they do not seem to have translated into commercial solutions or clinical trials. Even very recent papers [5, 6] provide almost no quantitative experiments on real data.

In order to alleviate the limitations of the current models, several studies have proposed to incorporate some machine learning components into the workflow, either to refine the model [4] or to detect uncertainties in the estimates [7, 6]. Yet, surprisingly no work so far has aimed at bypassing the whole speckle decorrelation model with a fully machine learning-based approach. This is probably due to the extreme difficulty of the problem, and the definition of meaningful image features. Recently though, deep learning approaches - and more particularly convolutional neural networks - have proven successful at solving even the most challenging image analysis problems [8].

In this paper, we hence investigate the use of deep learning for the estimation of relative motion between US images. We propose an end-to-end approach based on a convolutional neural network (CNN) that directly learns the relative 3D translations and rotations from a pair of images, and also suggest refinements to further improve the transformation estimates (Section 2). For the first time to our knowledge, we perform an extensive evaluation on 120 real US datasets including 100 acquired in clinical conditions (Section 3). Those experiments show that our method significantly outperforms the standard approaches and allows us to reconstruct long US sweeps with a very limited drift.

2 Methods

2.1 From speckle decorrelation to convolutional neural networks

Speckle patterns are seemingly random reflecting tissue inhomogeneities smaller than the ultrasound wavelength. Their partial correlation in successive US frames is exploited in the speckle decorrelation method as follows. The images are first divided into non-overlapping patches. Then the normalized cross-correlation between each patch of the first image and a set of patches from the second image in its neighborhood is computed. For every patch, the displacement that gives the best correlation is stored, which yields a 2D displacement map representing the in-plane motion. In order to retrieve the out-of-plane component of the local displacements, the maximum correlation value is used, which can be mapped to the elevational displacement using a statistical and physical model (see [2]

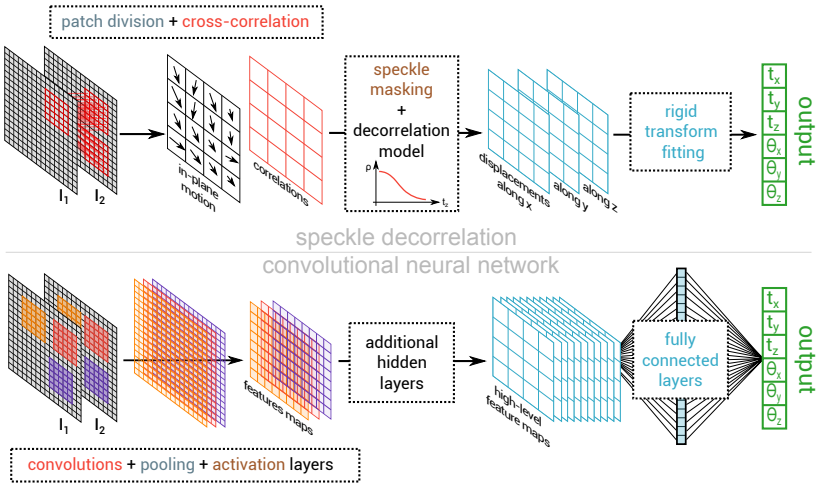

Fig. 1. Workflow comparison of speckle decorrelation (top) and convolutional neural network (bottom) for the estimation of the transformation parameters between two successive images. Related steps in the two approaches have the same color.

for instance). Unfortunately such models are only valid under Rayleigh scattering conditions, which means that only a subset of the patches - that also has to be automatically detected - may be used. Finally, a vector of parameters $\mathbf{p} = [t_x, t_y, t_z, \theta_x, \theta_y, \theta_z]^\top$ representing a rigid transformation $\mathbf{T}(\mathbf{p})$ with t and θ the translational and rotational components is fitted to the 3D vector field, usually with a robust algorithm in order to minimize the influence of outliers.

Trying to mimic this elaborate approach with a single CNN might seem overly ambitious or lead to uninterpretable results. Yet, as we show in Figure 1, it turns out that the two approaches do share some similarities. The analogy is far from perfect, but we believe that it gives some insight on why it makes sense to use a CNN. On the one hand, the basic steps of both approaches can be related: (*i*) the local cross-correlation operation may be approximated by a set of convolution filters, (*ii*) the patch-based approach that aggregates local information corresponds to the pooling layers of the network, (*iii*) the selection of reliable speckle features and areas in the image could be achieved via the activation layers. On the other hand, the more complex steps of the pipeline (the decorrelation model, the robust transformation fitting, etc.) are now replaced with a combination of non-linear operations whose modeling capabilities exceed all physical models but are more prone to overfitting. A strategy to alleviate this risk by adding simple and reliable prior information is proposed in Section 2.2.

We used a standard convolutional neural network architecture described in Figure 2. In all our experiments, the machine learning models are trained and tested using 2-fold patient cross validation for each dataset separately. Our algorithms are implemented in C++ and we used the Caffe framework for the deep

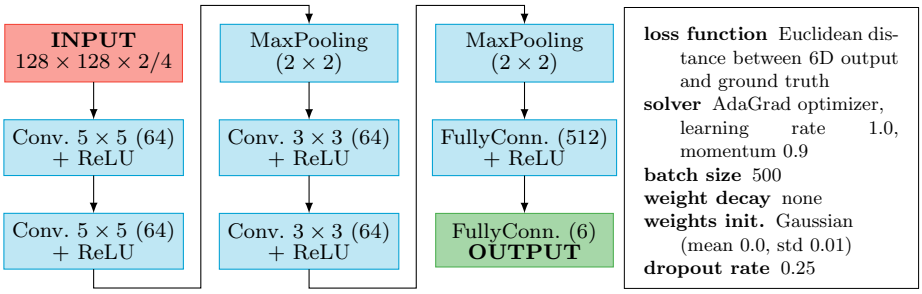

Fig. 2. Architecture of our convolutional neural networks and training parameters. All convolutions and pooling layers have a stride of 2 pixels.

learning components. Predicting the tracking of a whole sweep takes around 5 seconds on a standard computer with an NVIDIA GeForce GTX 1080 GPU.

2.2 Using optical flow as additional information

Even if neural networks are supposed to discover all necessary image features by themselves, the end-to-end problem that we are addressing remains quite challenging. One way of helping the network is to provide an estimate of the in-plane motion. While Dosovitskiy *et al.* have recently shown that neural networks are able to learn in-plane displacements [9], we hypothesize that by pre-computing an estimate of the in-plane displacement, we allow the neural network to focus on the most important part, namely the out-of-plane motion estimation.

We therefore compute a sub-pixel dense optical flow [1] and use this as additional channels of the images. Our network input now actually has 4 channels: the first two being the two successive images and the last two being the two components of the estimated vector field. Our experiments will show that this trick has a significant impact on the performance.

3 Experiments and results

Datasets acquisition and baseline methods. All sweeps used in our experiments were captured with a Cicada-64 research ultrasound machine by Cephasonics (Santa Clara, CA USA). We used a linear 128-element probe at 9MHz for generating the ultrasound images. The depth of all images was set to 5cm (with a focus at 2cm) and 256 scan-lines were captured per image. We used the B-mode images without any filtering or back-scan conversion, resampled with an isotropic resolution of 0.3 mm (this value was chosen to match the speckle scale, and we confirmed by cross-validation that this was indeed a suitable choice).

The probe was equipped with an optical target which was accurately tracked by a surgical system (Stryker Navigation System III). After thorough spatial and temporal image-to-sensor calibration, we were able to get a ground truth transformation with absolute positioning accuracy of around 0.2 mm according

to our tests. Since the ground truth has to be extremely precise from frame-to-frame, we also assured the temporal calibration exhibits neither jitter nor drift at all, thanks to the digital interface of the research US system and proper clock synchronization. Our experiments are based on three different datasets:

- a set of 20 US sweeps (7168 frames in total) acquired on a BluePhantom ultrasound biopsy phantom. The images contain mostly speckle but also a variety of masses that are either hyperechoic or hypoechoic;
- a set of 88 in-vivo tracked US sweeps (41869 frames in total) acquired on the forearms of 12 volunteers. Two different operators acquired at least three sweeps on both forearms of each participant;
- another 12 in-vivo tracked sweeps (6647 frames in total) acquired on the lower legs on a subset of the volunteers. This last set will be used to assess how the network generalizes to other anatomies.

The forearm and leg anatomy was chosen with the clinical application of peripheral vein mapping for bypass surgery or AV-fistula mapping in mind, which requires elongated sweeps to visualize vascular topology across a limb.

All sweeps have been acquired in a fixed direction (proximal to distal). This means that applying our algorithm on a reversed sweep would yield a mirrored result. However this limitation is not specific to our method, but is due to the problem in general being ill-posed. Besides, we believe that enforcing the acquisition direction of the sweeps is not a major constraint for the clinician.

We compared our algorithm to two baseline methods:

- a *linear motion*, which is the expected motion of the operator. This means that we set all parameters to their average value over all acquisitions: rotations and in-plane translations are almost zero while elevational translation t_z is constant around 2cm/s;
- the result of our implementation of a *speckle decorrelation* method: we filter each image to make the speckle pattern more visible as in [10], we divide each image in 15×15 patches and compute the corresponding patch-wise cross-correlations. We then use a standard exponential-based model to deduce the corresponding z-displacement from the correlation values (we were not able to fit more complex models). Finally we use RANSAC to compute a robust fit of the 6 transformation parameters to the displacement field.

Methods comparison. For each method and dataset, we compute error metrics on all transformation parameters but also in terms of final drift. Those numbers are reported in the first two tables of Figure 3 for the phantom acquisitions and the forearms dataset; the conclusions are similar for both datasets.

We first notice that assuming a perfectly *linear motion* gives the worst results of the four methods, which is mainly due to the out-of-plane translation t_z. This was expected since this component had the largest variability (it is easier for the operator to keep the US images parallel than to keep a constant speed). The *speckle decorrelation* approach does manage to significantly reduce all estimation errors by exploiting the correlations between the frames; nevertheless the out-of-plane error on t_z and therefore the overall drift is still quite high. On the

Table 1	avg. absolute error (mm/°)						final drift (mm)		
phantom dataset	t_x	t_y	t_z	θ_x	θ_y	θ_z	min	med.	max
linear motion	2.27	8.71	38.72	2.37	2.71	0.97	2.29	70.30	149.19
speckle decorrelation	4.96	2.21	29.89	2.10	4.46	1.93	12.67	47.27	134.93
standard CNN	2.25	5.67	14.37	2.13	1.86	0.98	14.31	26.17	65.10
CNN with optical flow	1.32	2.13	7.79	2.32	1.21	0.90	1.70	18.30	36.90

Table 2	avg. absolute error (mm/°)						final drift (mm)		
forearms dataset	t_x	t_y	t_z	θ_x	θ_y	θ_z	min	med.	max
linear motion	4.46	6.11	24.84	3.51	2.59	2.37	10.11	46.23	129.93
speckle decorrelation	4.36	4.09	18.78	2.53	3.02	5.23	9.19	36.36	98.95
standard CNN	6.30	5.97	6.15	2.82	2.78	2.40	3.72	25.16	63.26
CNN with optical flow	3.54	3.05	4.19	2.63	2.52	1.93	3.35	14.44	41.93
after speckle filtering	3.57	3.59	8.56	2.56	2.64	2.01	5.14	22.04	44.15

Table 3	avg. absolute error (mm/°)						final drift (mm)		
lower legs dataset	t_x	t_y	t_z	θ_x	θ_y	θ_z	min	med.	max
linear motion	4.49	4.84	39.81	4.39	2.18	2.46	37.35	73.40	143.42
speckle decorrelation	5.02	2.87	30.89	1.82	1.78	4.11	43.21	54.74	89.97
standard CNN	5.34	5.62	17.22	2.58	2.45	2.84	21.73	43.21	65.68
CNN with optical flow	4.14	3.91	17.12	1.94	2.58	2.15	25.79	40.56	52.72
CNN trained on legs	3.11	5.86	5.63	2.75	3.17	5.24	8.53	19.69	30.11

Fig. 3. Summary of the performance of the different methods on the three datasets. The parameter-wise errors are computed and averaged for every frame with respect to the first frame of the sweep. The final drift is defined as the distance between the last image center with the estimated tracking and ground truth.

other hand, the *standard CNN* without the optical flow channels is here able to produce results that are already better than the other approaches. One can notice though that the t_x and t_y errors are slightly higher than the speckle decorrelation method, especially on the forearm sweeps. Our guess is that the network focuses its effort on the t_z component because it represents the main part of the motion; learning the whole transformation more accurately would probably require a deeper network and a larger dataset. This can be fixed by adding the *optical flow* as input channels. We indeed see that t_x and t_y for instance are better estimated; the estimation of t_z is even further improved because the network can focus on the out-of-plane motions. In average, we observe on real clinical images a final drift of merely 1.45 cm over sequences longer than 20 cm, which is twice as accurate as speckle decorrelation. The hierarchy of the methods (linear $<$ speckle decorrelation $<$ standard CNN $<$ CNN with optical flow) was confirmed by paired signed-rank Wilcoxon tests which all yielded p-values lower than 10^{-6}.

In order to further demonstrate the efficiency of our method for out-of-plane estimation, we have recorded a separate sweep with a deliberately strongly varying speed and plotted the different predictions of the elevational translation in Figure 4. The first 100 and last 150 frames were recorded at an average speed of 0.3 mm/frame, while inbetween the speed has almost been doubled. Naturally,

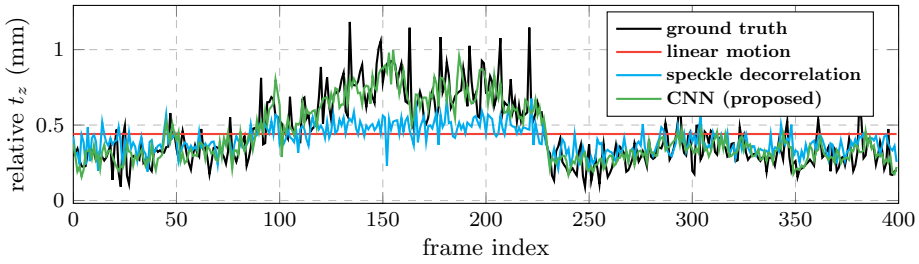

Fig. 4. Elevational translations t_z predicted values with different methods on an ultrasound sweep deliberately acquired with a strongly varying speed.

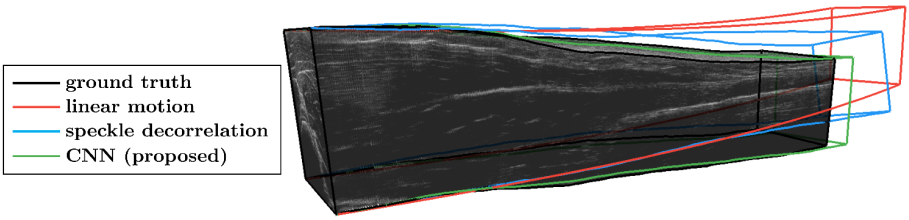

Fig. 5. Comparison of the trajectories reconstructed with different methods. This sample case corresponds to the median case in terms of estimation accuracy.

the *linear motion* method assumes a constant speed and will therefore yield major reconstruction artifacts. The *speckle decorrelation* approach does detect a speed change but strongly underestimates large motions. Only the *neural network* is able to follow the probe speed accurately. A qualitative comparison of the reconstructed trajectories on a sample sweep is also shown in Figure 5.

Influence of the noise filtering. In order to test the importance of the speckle noise, we compared the methods when applied on the images before and after applying the speckle filter built in the Cephasonics ultrasound system. As we can see in the last row of Table 2, learning and testing on the unfiltered images yields better tracking estimation. Speckle patterns are therefore important for the neural network, in particular for the estimation of the out of plane translation. This result therefore tends to validate the intuition of the research community that speckle is indeed important, but not necessary since using the CNN on filtered images already gives better results than the other methods.

Generalization to other anatomies. Another interesting question is to assess how well such a network can generalize to other applications: does it really learn the motion from general statistics, or does it overfit to some anatomical structures present in the image? The results reported in Table 3 show a significant degradation of the results for all methods (since they all have been calibrated and learned on the forearms dataset). The in-plane displacements are still recovered with a similar accuracy but the error on the out-of-plane translation t_z has strongly increased. However, we can notice that our CNN-based method still generalizes better than the others to a new kind of images. This preliminary ex-

periment shows that the accuracy is strongly dependent on the target anatomy but gives hope regarding the capabilities of our network. For comparison, we also report the accuracy obtained with a CNN trained on this specific dataset, which is only slightly worse than on forearms (due to the smaller dataset size).

4 Conclusion

This paper introduced a sensorless 3D ultrasound system with a tracking estimation based on deep learning. We showed how CNNs relate to the standard method of speckle decorrelation but offer a much stronger complexity that is able to learn the relationship between speckle and out-of-plane motion. Our evaluation, the first one on such a large dataset, showed great results (7% drift wrt. the sweep length) for peripheral vein mapping.

We believe that our work paves the way for many further clinical applications where reconstructing 3D volumes from standard 2D ultrasound clips may be valuable. The reconstruction error may then also be further reduced by restricting the imaging protocol or adding redundant information from perpendicular clips or panoramic stitched data to use during 3D pose estimation. It would also be interesting to investigate the dependency on ultrasound system parameters (probe, depth, frequency, etc). Last but not least, we also plan to try more complex network like recurrent neural networks.

References

1. Farnebäck, G.: Two-frame motion estimation based on polynomial expansion. In: Scandinavian conference on Image analysis, Springer (2003) 363–370
2. Prager, R.W., Gee, A.H., Treece, G.M., Cash, C.J., Berman, L.H.: Sensorless freehand 3-d ultrasound using regression of the echo intensity. Ultrasound in medicine & biology **29**(3) (2003) 437–446
3. Gee, A.H., Housden, R.J., Hassenpflug, P., Treece, G.M., Prager, R.W.: Sensorless freehand 3d ultrasound in real tissue: speckle decorrelation without fully developed speckle. Medical image analysis **10**(2) (2006) 137–149
4. Laporte, C., Arbel, T.: Learning to estimate out-of-plane motion in ultrasound imagery of real tissue. Medical image analysis **15**(2) (2011) 202–213
5. Gao, H., Huang, Q., Xu, X., Li, X.: Wireless and sensorless 3D ultrasound imaging. Neurocomput. **195**(C) (June 2016) 159–171
6. Tetrel, L., Chebrek, H., Laporte, C. In: Learning for Graph-Based Sensorless Freehand 3D Ultrasound. Springer International Publishing (2016) 205–212
7. Conrath, J., Laporte, C.: Towards improving the accuracy of sensorless freehand 3d ultrasound by learning. In: International Workshop on Machine Learning in Medical Imaging, Springer (2012) 78–85
8. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521** (2015) 436–444
9. Dosovitskiy, A., Fischer, P., Ilg, E., Hausser, P., Hazirbas, C., Golkov, V., v.d. Smagt, P., Cremers, D., Brox, T.: Flownet: Learning optical flow with convolutional networks. In: IEEE International Conference on Computer Vision (ICCV)
10. Afsham, N., Rasoulian, A., Najafi, M., Abolmaesumi, P., Rohling, R.: Nonlocal means filter-based speckle tracking. IEEE transactions on ultrasonics, ferroelectrics, and frequency control **62**(8) (2015) 1501–1515