Localizing 2D Ultrasound Probe from Ultrasound Image Sequences Using Deep Learning for Volume Reconstruction

Kanta Miura[1], Koichi Ito[1], Takafumi Aoki[1], Jun Ohmiya[2], and Satoshi Kondo[2]

[1] Graduate School of Information Sciences, Tohoku University,
6-6-05, Aramaki Aza Aoba, Aoba-ku, Sendai-shi, Miyagi, 9808579, Japan.
kanta@aoki.ecei.tohoku.ac.jp
[2] AI Technology Development Division IoT Service Platform Development Operations, Konica Minolta, Inc.,
1-2, Sakura-machi, Takatsuki-shi, Osaka, 5698503, Japan.

Abstract. This paper presents an ultrasound (US) volume reconstruction method only from US image sequences using deep learning. The proposed method employs the convolutional neural network (CNN) to estimate the position of a 2D US probe only from US images. Our CNN model consists of two networks: feature extraction and motion estimation. We also introduce the consistency loss function to enforce. Through a set of experiments using US image sequence datasets with ground-truth motion measured by a motion capture system, we demonstrate that the proposed method exhibits the efficient performance on probe localization and volume reconstruction compared with the conventional method.

Keywords: ultrasound · volume reconstruction · CNN · probe localization.

1 Introduction

Ultrasound (US) imaging has a number of advantages in medical diagnosis such as high spatial resolution, real-time imaging, and non-invasiveness. Recently, three-dimensional (3D) US [13] has attracted much attention as a valuable imaging tool for a diagnostic procedure because of the above advantages of US. If 3D US can be acquired using only the current US system with a 2D US probe, 3D US may be allowed to be used in place of other imaging modalities such as CT, MRI or PET, e.g., the point of care in an emergency situation requiring the rapid diagnosis, muscle and blood analysis in sports medicine, etc. Among 3D US acquisition protocols, we focus on the freehand protocol [4] because of its cost-effectiveness and flexibility. 3D volume data can be reconstructed from a 2D US image sequence by integrating a set of US images according to the position of the US prove. The quality of 3D volume data significantly depends on the accuracy of probe localization, i.e., 3D motion estimation of a 2D US probe in the acquisition protocol with freehand scanning.

The initial approach of localizing a 2D US probe is to use the special devices such as an electromagnetic tracking device [17, 7] and an optical tracker [6, 18]. The accuracy of probe localization is high, while such special devices require much cost and may sacrifice smooth scanning. The simple approach is to use markers to estimate the motion of a 2D US probe [9, 16, 12]. The cost of the system is cheap, while markers are attached on the skin surface, resulting in decreasing the flexibility and acceptability. Another approach is to use a camera, which is mounted on a 2D US probe. The motion of the probe is estimated from a video sequence of skin patterns captured by a camera using simultaneous localization and mapping (SLAM) [19] or structure from motion (SfM) [11, 10]. This approach is cost-effective, while a camera may be intrusive for an operator. The challenging task is to estimate the motion of a US probe only from a US image sequence. Balakrinshnan et al. [1] proposed a similarity metric, which computes the similarity between two consecutive US images by correlating the parametric representation of image texture, to estimate out-of-plane motion in US probe sweeping. Prevost et al. [15] proposed a 2D US probe localization method using a convolutional neural network (CNN), which estimates the motion of a 2D US probe by image-based tracking. This method learns the relative 3D translations and rotations from a pair of images with additional information of optical flow, which is used to improve the accuracy of motion estimation. The CNN architecture of this method is relatively simple, which consists of 4 convolution layers, 2 pooling layers, and 2 fully-connected layers. Their latest work in [14] also used an inertial measurement unit (IMU), which was mounted on a US prove, to improve the accuracy of estimating 3D rotation.

In this paper, we propose a 2D US probe localization method only from US image sequences using deep learning. We consider a new CNN architecture for estimating the motion between two US images inspired by Prevost's work [15, 14]. Our CNN architecture includes motion features obtained from FlowNetS [2]. We introduce the consistency loss function to improve the accuracy of motion estimation. We create a large-scale dataset of US image sequences with the ground-truth probe motion for evaluating the methods. The US image sequences are acquired by scanning forearm, breast phantom and hypogastric phantom, where the number of images of each target is 30,801, 8,940, and 6,242, respectively. The contribution of this work is summarized as follows:

1. propose a new CNN architecture for localizing a 2D US probe for volume reconstruction and
2. introduce a consistency loss function to improve the accuracy of probe localization.

2 Methods

This section describes our CNN architecture for estimating the motion between two US images and its loss functions to improve the accuracy of motion estimation.

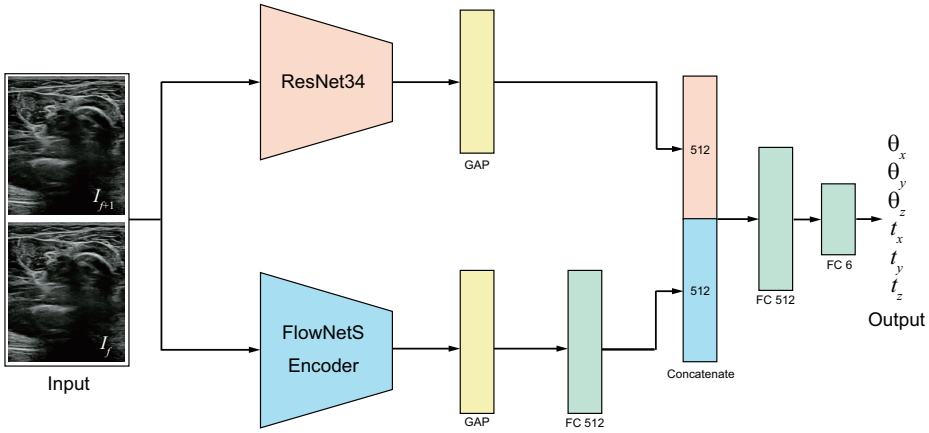

Fig. 1. Network architecture of our proposed CNN.

2.1 Network Architecture

In the previous work by Prevost et al. [15, 14], they proposed a simple CNN architecture for estimating the motion between two US images, which consists of 4 convolution layers, 2 pooling layers, and 2 fully-connected layers. They used 4-channel input, which consists of the two US images and the two components of the vector field estimated by optical flow estimation [3]. The optical flow between the two US images is not always accurately estimated by [3] from our empirical observation. Fig. 1 shows the network architecture of our proposed CNN for estimating the motion between two US images. This architecture consists of localization and optical flow estimation networks. In this paper, we employ ResNet34 [8] for localization network and the encoder of FlowNetS [2] for optical flow estimation network. The feature vector extracted from ResNet34 is reduced to a 512-dimensional feature vector by Global Average Pooling (GAP). The feature vector extracted from FlowNetS is also reduced to a 512-dimensional feature vector by GAP and the fully-connected layer. Then, two feature vectors are concatenated before the last two fully-connected layers. The output of CNN is 6 parameters consisting of 3 rotation angles $(\theta_x, \theta_y, \theta_z)$ and 3 translations (t_x, t_y, t_z), where $\boldsymbol{p} = \{\theta_x, \theta_y, \theta_z, t_x, t_y, t_z\}$. We employ FlowNetS pre-trained by the Flying Chairs dataset[3] and all the weight parameters are fixed in both training and test.

2.2 Loss function

We employ the loss function defined by the Euclidean distance between estimated 6 parameters and the ground truth as well as the previous work [15, 14], which

[3] `https://lmb.informatik.uni-freiburg.de/resources/datasets/FlyingChairs.en.html`

is given by

$$L_{\mathrm{Euc}} = ||\boldsymbol{P}^g - \hat{\boldsymbol{P}}||_2, \tag{1}$$

where $\boldsymbol{P}^g$ indicates the ground-truth vector of parameters and $\hat{\boldsymbol{P}}$ indicates the estimated vector. We also consider introducing the new loss function to improve the accuracy of the motion between US image frames. Let the rotation and the translation from image frame I_f to I_{f+1} be $\boldsymbol{R}_{f\to f+1}$ and $\boldsymbol{t}_{f\to f+1}$, and their inverses be $\boldsymbol{R}_{f+1\to f}$ and $\boldsymbol{t}_{f+1\to f}$. $\boldsymbol{R}_{f\to f+1}$ and $\boldsymbol{t}_{f\to f+1}$ are estimated from I_f and I_{f+1} by CNN, and $\boldsymbol{R}_{f+1\to f}$ and $\boldsymbol{t}_{f+1\to f}$ are also estimated by CNN when reversing the order of the inputs. A point on I_f should be reprojected onto the same position when applying the transformation $[\boldsymbol{R}_{f\to f+1}|\boldsymbol{t}_{f\to f+1}]$ and then $[\boldsymbol{R}_{f+1\to f}|\boldsymbol{t}_{f+1\to f}]$. This is known as the reprojection error in stereo vision, and we can apply the similar idea of the left-right consistency loss in stereo vision [5] to our method. Let a point on the image frame I_f be $\boldsymbol{P}_f$ and a point reprojected from the consecutive image frame I_{f+1} be $\boldsymbol{P}'_f$, respectively. The point $\boldsymbol{P}'_f$ is calculated using the rotation and translation between the two image frames as follows:

$$\boldsymbol{P}'_f = \boldsymbol{R}_{f+1\to f}(\boldsymbol{R}_{f\to f+1}\boldsymbol{P}_{f1} + \boldsymbol{t}_{f\to f+1}) + \boldsymbol{t}_{f+1\to f}, \tag{2}$$

where We consider the following consistency loss function:

$$L_{\mathrm{Cons}} = ||\boldsymbol{P}_f - \boldsymbol{P}'_f||_2. \tag{3}$$

3 Materials

We create a large-scale dataset of US image sequences with the ground-truth probe motion for evaluating the methods. The target objects are forearm of 5 subjects, breast phantom and hypogastric phantom in the dataset. US image sequences are acquired by SONIMAGE HS1 (Konica Minolta, Inc.) with L18-4 linear probe (center frequency: 10MHz) for forearm and breast phantom and with C5-2 convex probe (center frequency: 3.5MHz) for hypogastric phantom, where the field of view (FOV) of US images is 40×38mm, the frame rate is 30fps, the recording time is about 6 seconds (180 frames), and the size of each US image frame is 442 × 526 pixels. The number of scans (image frames) is 190 (30,801) for forearm, 60 (8,940) for breast phantom, and 40 (6,242) for hypogastric phantom. The ground-truth position of the US probe is measured by V120:Trio (OptiTrack), where 5 markers are attached on the US probe to capture its motion.

4 Experiments

In the experiments, we separate the dataset into training, validation, and test data, where the training data is 180 scans (27,948 image frames) from forearm of 2 subjects and two phantoms, the validation data is 30 scans (5,176 image frames) from forearm of 1 subject, and the test data is 80 scans (12,859 frames) from forearm of 2 subjects. Each image frame with 442 × 526 pixels is cropped

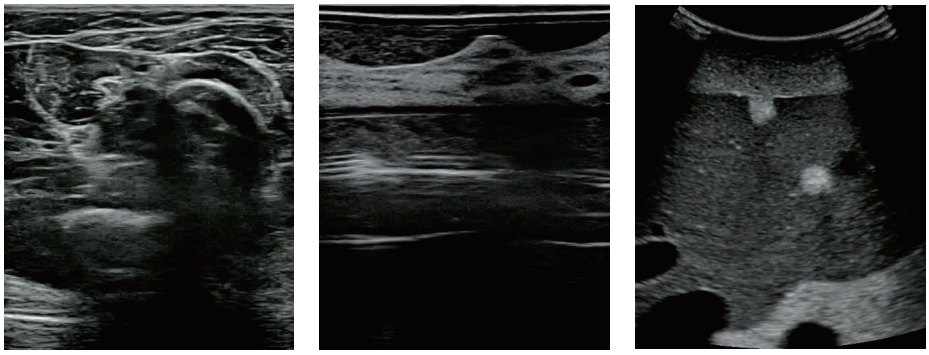

Fig. 2. Example of the US image frame acquired from (a) forearm, (b) breast phantom, and (c) hypogastric phantom.

the center region with 442×442 pixels, and then is resized to 256×256 pixels. The pixel value of each resized image is normalized to have zero mean and the unit variance.

The training parameters of our method are as follows: the optimizer is AdaGrad, the learning rate is 1e-3, the batch size is 64, the number of epochs is 30, and 25% dropout is added after the fully-connected layers except the last one. All the methods are implemented using PyTorch 1.0.0 on Intel(R) Xeon(R) W-2133 CPU 3.60GHz with GeForce RTX 2080 Ti. We evaluate the accuracy of each parameter estimated by the conventional method [15] and our methods using mean absolute error (MAE), where we consider 4 combinations for the proposed method in the following ablation study. Note that we implemented the conventional method according to the paper [15] since an official implementation is not provided. The conventional method was trained and evaluated under the same experimental condition.

Table 1 shows the summary of the ablation study. There is no significant difference in estimation accuracy depending on the network architecture comparing the first and second rows of Table 1. The estimation accuracy is comparable when adding FlowNetS to the proposed method (i) comparing the second and third rows of Table 1. The estimation accuracy is improved when adding the consistency loss function to the proposed method (i) comparing the first and third rows of Table 1 The combination of loss functions can limit the search space of parameter optimization in CNN. The proposed method (iv), which employs all the techniques, exhibits the best estimation accuracy in the methods except for t_z as observed in the forth row of Table 1. Fig. 3 shows the temporal variation of parameters estimated by each method. The conventional method cannot estimate large motion and therefore show the average temporal variation. The proposed method (i) shows a temporal variation close to the ground truth, while it may deviate significantly. The proposed method (iv) shows similar temporal variation to the ground truth for all the parameters except for t_z.

Table 1. Summary of the ablation study (OF: Optical flow, FN: FlowNetS).

Method		OF	FN	L_{Euc}	L_{Cons}	MAE(degree/mm)					
						θ_x	θ_y	θ_z	t_x	t_y	t_z
Prevost et al. [14]		✓		✓		0.58	1.28	0.49	0.69	0.16	0.77
Ours	(i)	✓		✓		0.60	1.26	0.52	0.72	0.18	**0.76**
	(ii)	✓	✓	✓		0.61	1.28	0.52	0.74	0.18	0.78
	(iii)	✓		✓	✓	0.56	1.23	0.47	0.66	**0.15**	0.82
	(iv)	✓	✓	✓	✓	**0.53**	**1.21**	**0.47**	**0.64**	**0.15**	0.80

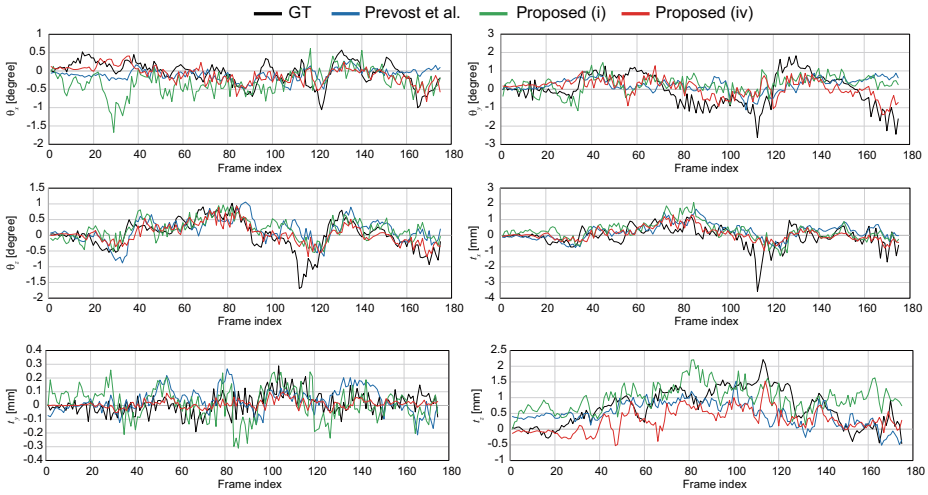

Fig. 3. Temporal variation of parameters estimated by each method.

Fig. 4 shows the reconstructed US volume data using the probe location of the ground truth, the conventional method, and the proposed methods (i)∼(iv). Each volume data is reconstructed using StradView[4]. The conventional method cannot handle the large motion of the US probe since the estimated motion is similar to the linear motion. Although the proposed methods (i) and (ii) attempt to estimate a large motion of the US probe, the estimated motion is rather large. The proposed methods (iii) and (iv) exhibit better performance than other methods since the shape of the reconstructed volume is similar to that of the ground truth.

5 Conclusion

In this paper, we proposed a 2D US probe localization method only from US image sequences using deep learning. Our CNN architecture extracts texture features and motion features, and estimate the motion between two US image

[4] `https://mi.eng.cam.ac.uk/Main/StradView`

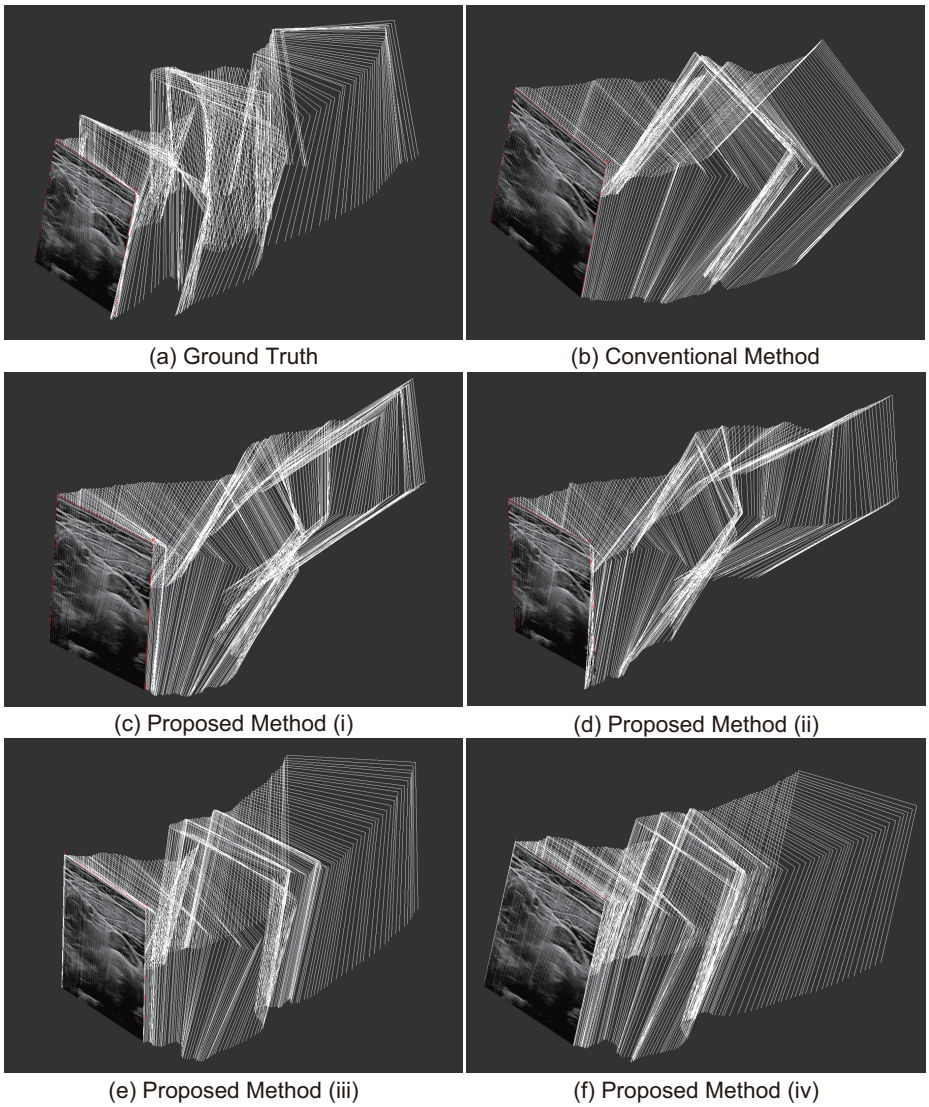

Fig. 4. Example of reconstructed US volume data: (a) ground truth, (b) conventional method, and (c)∼(f) proposed method (i)∼(iv).

frames. We considered the combination of loss functions to improve the accuracy of motion estimation. Through a set of experiments using our dataset of forearm, breast phantom, and hypogastric phantom, we demonstrated that our method exhibited better accuracy of probe localization than the conventional method. In future work, we will develop a 2D US probe with a small camera to support a

large variety of probe motion to realize a free-hand 3D US reconstruction system for practical use.

References

1. Balakrishnan, S., Patel, R., Illanes, A., Friebe, M.: Novel similarity metric for image-based out-of-plane motion estimation in 3D ultrasound. Proc. Int'l Conf. IEEE Engineering in Medicine and Biology Society pp. 5739–5742 (Jul 2019)
2. Dosovitskiy, A., Fischer, P., Ilg, E., Häusser, P., Hazirbas, C., Golkov, V., Smagt, P., Cremers, D., Brox, T.: Flownet: Learning optical flow with convolutional networks. Proc. IEEE Int'l Conf. Computer Vision pp. 2758–2766 (Dec 2015)
3. Farnebäck, G.: Two-frame motion estimation based on polynomial expansion. LNCS 2749 (SCIA 2003) pp. 363–370 (2003)
4. Gee, A., Prager, R., Treece, G., Berman, L.: Engineering a freehand 3D ultrasound system. Pattern Recognition Letters **24**(4–5), 757–777 (Feb 2003)
5. Godard, C., Aodha, O.M., Brostow, G.J.: Unsupervised monocular depth estimation with left-right consistency. Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition pp. 270–279 (Jul 2017)
6. Goldsmith, A., Pedersen, P., Szabo, T.: An inertial-optical tracking system for portable, quantitative, 3D ultrasound. Proc. IEEE Int'l Ultrasonics Symp. pp. 45–49 (Nov 2008)
7. Hastenteufel, M., Vetter, M., Meinzer, H.P., Wolf, I.: Effect of 3D ultrasound probes on the accuracy of electromagnetic tracking systems. Ultrasound in Med. & Biol. **32**(9), 1359–1368 (Sep 2006)
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition pp. 770–778 (Jun 2016)
9. Horvath, S., Galeotti, J., Wang, B., Perich, M., Wang, J., Siegel, M., Vescovi, P., Stetten, G.: Towards an ultrasound probe with vision: Structured light to determine surface orientation. LNCS 7264 (AE-CAI 2011) pp. 58–64 (Sep 2012)
10. Ito, K., Yodokawa, K., Aoki, T., Ohmiya, J., Kondo, S.: A probe-camera system for 3D ultrasound image reconstruction. LNCS 10549 (POCUS 2017) pp. 129–137 (Sep 2017)
11. Ito, S., Ito, K., Aoki, T., Ohmiya, J., Kondo, S.: Probe localization using structure from motion for 3D ultrasound image reconstruction. Proc. Int'l Symp. Biomedical Imaging pp. 68–71 (Apr 2017)
12. Lange, T., Kraft, S., Eulenstein, S., Lamecker, H., Schlag, P.: Automatic calibration of 3D ultrasound probes. Proc. Bildverarbeitung für die Medizin pp. 169–173 (Mar 2011)
13. Nelson, T.R., Pretorius, D.H.: Three-dimensional ultrasound imaging. Ultrasound in Medicine & Biology **24**(9), 1243–1270 (Dec 1998)
14. Prevost, R., Salehi, M., Jagoda, S., Kumar, N., Sprung, J., Ladikos, A., Bauer, R., Zettinig, O., Wein, W.: 3D freehand ultrasound without external tracking using deep learning. Medical Image Analysis **48**, 187–202 (Aug 2018)
15. Prevost, R., Salehi, M., Sprung, J., Ladikos, A., Bauer, R., Wein, W.: Deep learning for sensorless 3D freehand ultrasound imaging. LNCS 8674 (MICCAI 2017) pp. 628–636 (Sep 2017)
16. Rafii-Tari, H., Abolmaesumi, P., Rohling, R.: Panorama ultrasound for guiding epidural anesthesia: A feasibility study. LNCS 6689 (IPCAI 2011) pp. 179–189 (Jun 2011)

17. Rousseau, F., Hellier, P., Barillot, C.: A fully automatic calibration procedure for freehand 3D ultrasound. Proc. IEEE Int'l Symp. Biomedical Imaging pp. 985–988 (Jul 2002)
18. Stolka, P., Kang, H., Choti, M., Boctor, E.: Multi-DoF probe trajectory reconstruction with local sensors for 2D-to-3D ultrasound. Proc. IEEE Int'l Symp. Biomedical Imaging pp. 316–319 (Apr 2010)
19. Sun, S.Y., Gilbertson, M., Anthony, B.: Probe localization for freehand 3D ultrasound by tracking skin features. LNCS 8674 (MICCAI 2014) pp. 365–372 (Sep 2014)