

# Probe Localization for Freehand 3D Ultrasound by Tracking Skin Features

Shih-Yu Sun, Matthew Gilbertson, and Brian W. Anthony

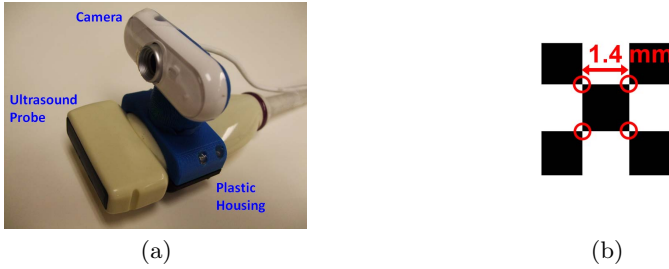
Laboratory for Manufacturing and Productivity,  
Medical Electronic Device Realization Center,  
Massachusetts Institute of Technology, Cambridge, MA, USA

**Abstract.** Ultrasound probe localization with respect to the patient's body is essential for freehand three-dimensional ultrasound and image-guided intervention. However, current methods for probe localization generally involve bulky and expensive equipment. In this paper, a highly cost-effective and miniature-mobile system is described for 6-DoF probe localization that is robust to rigid patient motion. In this system, skin features in the scan region are recorded at each ultrasound scan acquisition by a lightweight camera rigidly mounted to the probe. A skin map is built based on the skin features and optimal probe poses are estimated in a Bayesian probabilistic framework that incorporates a prior motion model, camera frames, and ultrasound scans. Through freehand scanning on three different body parts, it is shown that on average, for every probe travel distance of 10 mm, the translational and rotational errors are  $0.91 \pm 0.49$  mm and  $0.55^\circ \pm 0.17^\circ$ , respectively. The 3D reconstructions were also validated by comparison with real ultrasound scans.

## 1 Introduction

Localization of the ultrasound probe with respect to the patient's body in six degrees of freedom (6 DoF) is essential for freehand three-dimensional ultrasound (3D US) and image-guided intervention. Currently freehand probe motion is typically tracked by using an optical or electromagnetic (EM) tracker, which is able to provide sub-millimeter accuracy in real time. However, these tracking methods often require expensive and bulky equipment. Additionally, an optical tracker requires a direct line of sight between the optical system and markers attached to the probe, and an EM tracker is sensitive to ferromagnetic materials, both of which pose challenges to the clinical use of these methods.

Inexpensive methods for probe localization have been investigated, sometimes with the aid of inertial sensing. In [5], structured light is used to estimate probe orientation against the skin surface. [4] and [11] describe the use of systems similar to optical mice to track skin features from a small distance for determining 2D probe translation along the skin surface. [10] describes probe tracking by affixing a specialized strip with high-contrast markers to the scan region, which is suitable for pre-determined scan paths. These methods are robust to rigid patient motion since probe poses are determined with respect to the patient's body,



**Fig. 1.** (a) The US probe with a rigidly mounted camera for recording skin features during a freehand US scan. (b) A square sticker with known dimensions that is affixed to the skin surface of the scan region for scale calibration.

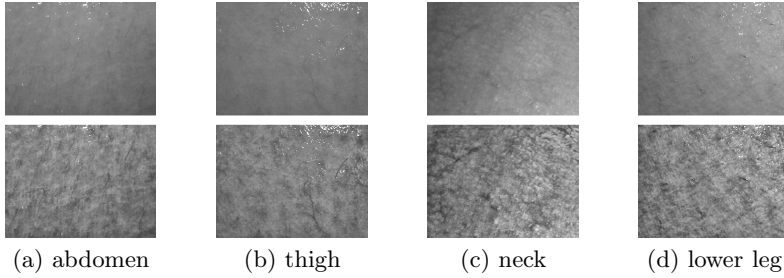
and not an independent coordinate system as in an optical or EM tracker. This robustness could also be achieved by tracking markers on both the probe and patient’s body simultaneously using a single optical system [2]. Finally, probe tracking based on only speckle decorrelation is possible [3].

Previously, we have developed a cost-effective and miniature-mobile system for 6-DoF US probe tracking that is robust to rigid patient motion [12]. A lightweight camera is mounted to the probe to record artificial skin features when acquiring 2D US scans. From the video of skin features, a skin map is built and the probe pose corresponding to each camera frame is estimated. In this current paper, our previous work is extended in two aspects. First, the probe is tracked by using natural skin features instead of artificial features. To our knowledge, this work is the first to demonstrate the use of human skin features as body landmarks. Second, a Bayesian framework is formulated to fuse information from a motion model, camera frames, and US scans for optimal pose estimation. The system performance was evaluated on three body parts.

## 2 Methods

### 2.1 Hardware

A lightweight, low-cost camera (Macally IceCam2) is mounted to a linear array US probe for tracking skin features, as shown in Fig. 1(a). The camera is focused at the skin surface from about 27 mm above, and has a  $640 \times 480$  field of view (FOV), which maps to around  $28 \times 21$  mm on the skin. Camera intrinsics were estimated by standard calibration. Transformation of coordinates from the US scan to the camera was estimated by using the single-wall method [9]. Camera frames and 2D US scans were acquired synchronously at around 10 frames per second using the Terason t3000 ultrasound scanning system. The square sticker with known dimensions shown in Fig. 1(b) was affixed to the skin surface of the scan region and the four corners (marked by red circles) were manually selected in two camera frames for scale calibration in visual SLAM (Section 2.3).



**Fig. 2.** Natural skin features covered by evenly applied transmission gel at four different body parts before (top) and after CLAHE (bottom). These features were recorded from about 27 mm above using the lightweight camera shown in Fig. 1(a).

## 2.2 Enhanced Natural Skin Features

Performance of our system hinges on robust tracking of natural skin features. It was found that, using the lightweight camera at a short distance to the skin surface, various skin features could be observed, including skin texture, hair, pores, and uneven skin tone. These features, however, become less observable when the skin surface is covered by transmission gel. In order to increase the number of skin features that could be robustly tracked, the recorded camera images are enhanced by contrast limited adaptive histogram equalization (CLAHE) before feature detection. Fig. 2 shows natural skin features at four different body parts of a human subject, which are more observable after CLAHE.

## 2.3 Probe Localization in a Bayesian Probabilistic Framework

The skin features recorded in freehand scanning are used to map the skin surface and estimate the 6-DoF camera pose corresponding to each 2D US scan, which is a process often called visual SLAM (Simultaneous Localization and Mapping). A Bayesian framework is formulated to fuse the information from visual SLAM with a prior motion model and US scan regularity for optimal pose estimation.

**Visual SLAM.** Visual SLAM is initialized by two-frame stereo. SIFT feature correspondences between the two frames are first found [7]. Based on the correspondences, the relative pose between the two frames is robustly estimated by applying the five-point algorithm within a RANSAC scheme [8]. The 3D skin map points are then computed by triangulating the feature correspondences. These two initial frames form the set of “keyframes” for future map extension.

Subsequent camera poses are estimated based on the initial map. A constant-velocity motion model is employed to predict the current camera pose based on previous estimates. Correspondences between the map points and features in the current frame are established by windowed searches for the most similar SIFT descriptors based on the predicted pose. The current pose is then robustly estimated by applying the EPnP algorithm within the RANSAC scheme [6].

To maintain the number of map points in the camera FOV, new points are added to the skin map when the tracked points cover less than one-third of the current FOV. SIFT feature correspondences are established between the current frame and the closest keyframe. Correspondences that agree with the current pose estimate are triangulated and these 3D map points are added to the skin map. The current frame is added to the set of keyframes for future map extension. Quality of visual SLAM is judged by the reprojection errors as described below.

**Bayesian Framework.** For optimal pose estimation, a Bayesian framework is formulated to fuse information from visual SLAM, US scans, and a prior motion model. The 6-DoF camera pose at time instance  $i$  is denoted by the vector  $\mathbf{v}_i$ , which includes 3-DoF translation and 3-DoF rotation.  $\dot{\mathbf{v}}_i$  denotes the temporal derivative of  $\mathbf{v}_i$ . Further,  $I_i$  denotes US scan intensities at  $i$ ,  $\mathcal{U}_i$  the set of features tracked in camera frame  $i$ , and  $\mathcal{X}$  the set of 3D map points. Assuming Markov conditional independence, the posterior probability of  $\mathbf{v}_i$  is written as:

$$P(\mathbf{v}_i | \mathbf{v}_{i-1}, \dot{\mathbf{v}}_{i-1}, I_i, I_{i-1}, \mathcal{X}, \mathcal{U}_i) \propto P(\mathbf{v}_i | \mathbf{v}_{i-1}, \dot{\mathbf{v}}_{i-1}, I_{i-1}, \mathcal{X}) P(\mathcal{U}_i, I_i | \mathbf{v}_{i-1}, \dot{\mathbf{v}}_{i-1}, I_{i-1}, \mathcal{X}, \mathbf{v}_i). \quad (1)$$

This formula can be simplified by conditional independence. In the first term, conditioned on  $\mathbf{v}_{i-1}$  and  $\dot{\mathbf{v}}_{i-1}$ ,  $\mathbf{v}_i$  depends only on instantaneous acceleration and is thus independent of both  $I_{i-1}$  and  $\mathcal{X}$ . Similarly, conditioned on  $I_{i-1}$ ,  $\mathbf{v}_{i-1}$  and  $\mathbf{v}_i$ ,  $I_i$  is independent of  $\mathcal{U}_i$ ,  $\mathcal{X}$  and  $\dot{\mathbf{v}}_{i-1}$ . Conditioned on  $\mathbf{v}_i$  and  $\mathcal{X}$ ,  $\mathcal{U}_i$  is independent of  $I_{i-1}$ ,  $I_i$ ,  $\mathbf{v}_{i-1}$  and  $\dot{\mathbf{v}}_{i-1}$ . Hence, the optimal estimate  $\mathbf{v}_i^*$  satisfies:

$$\mathbf{v}_i^* = \underset{\mathbf{v}_i}{\operatorname{argmax}} P(\mathbf{v}_i | \mathbf{v}_{i-1}, \dot{\mathbf{v}}_{i-1}) P(\mathcal{U}_i | \mathcal{X}, \mathbf{v}_i) P(I_i | \mathbf{v}_{i-1}, \mathbf{v}_i, I_{i-1}). \quad (2)$$

Modeling of the three information sources is described in the following.

*Prior Motion Model.* We assume a constant-velocity model and change in velocity  $\dot{\mathbf{v}}_i$  is modeled by mutually independent zero-mean Gaussian random variables, where the translational and rotational variances are  $\sigma_t^2$  and  $\sigma_R^2$ , respectively. Approximating  $\dot{\mathbf{v}}_i$  by  $(\mathbf{v}_i - \mathbf{v}_{i-1})$  and denoting the  $k$ -th element of  $\dot{\mathbf{v}}_i$  by  $\dot{\mathbf{v}}_{i,k}$ , the estimate  $\mathbf{v}_i$  that maximizes  $P(\mathbf{v}_i | \mathbf{v}_{i-1}, \dot{\mathbf{v}}_{i-1})$  also minimizes the energy  $E_{\text{prior}}(\mathbf{v}_i)$ :

$$E_{\text{prior}}(\mathbf{v}_i) = \frac{1}{\sigma_t^2} \sum_{k=1}^3 (\dot{\mathbf{v}}_{i,k} - \dot{\mathbf{v}}_{i-1,k})^2 + \frac{1}{\sigma_R^2} \sum_{k=4}^6 (\dot{\mathbf{v}}_{i,k} - \dot{\mathbf{v}}_{i-1,k})^2. \quad (3)$$

*Reprojection Errors in Visual SLAM.* Given  $\mathbf{v}_i$  and tracked map points  $\mathcal{X}_i \subseteq \mathcal{X}$ , image coordinates of tracked feature points  $\mathbf{u}_{i,k} \in \mathcal{U}_i$  are modeled by mutually independent Gaussian random variables, the means being projections of the corresponding map points  $\mathbf{X}_{i,k} \in \mathcal{X}_i$ . Denoting the projection of  $\mathbf{X}_{i,k}$  to a camera with pose  $\mathbf{v}_i$  by  $\text{Proj}(\mathbf{X}_{i,k}, \mathbf{v}_i)$  and the number of tracked points by  $N_i$ , the estimate  $\mathbf{v}_i$  that maximizes  $P(\mathcal{U}_i | \mathcal{X}, \mathbf{v}_i)$  minimizes the energy  $E_{\text{reproj}}(\mathbf{v}_i)$ :

$$E_{\text{reproj}}(\mathbf{v}_i) = \frac{1}{\sigma_u^2} \sum_{k=1}^{N_i} (\mathbf{u}_{i,k} - \text{Proj}(\mathbf{X}_{i,k}, \mathbf{v}_i))^2. \quad (4)$$

*Ultrasound Scan Regularity.* When two closely spaced US scans are accurately localized, intensities of their neighboring pixels should be similar. Denoting the  $k$ -th pixel intensity in scan  $I_i$  by  $I_{i,k}$  and its image coordinates by  $\mathbf{p}_{i,k}$ , the function  $F(\mathbf{p}_{i,k}, \mathbf{v}_{i-1}, \mathbf{v}_i, I_{i-1})$  gives its corresponding pixel intensity in scan  $I_{i-1}$ , which involves projection of  $\mathbf{p}_{i,k}$  onto the plane of  $I_{i-1}$  and interpolation on intensities. The sum of absolute intensity differences between all the corresponding pixels is modeled by an exponential random variable. Hence, with  $M$  pixels in each scan,  $\mathbf{v}_i$  that maximizes  $P(I_i | \mathbf{v}_{i-1}, \mathbf{v}_i, I_{i-1})$  also minimizes the energy  $E_{\text{scan}}(\mathbf{v}_i)$ :

$$E_{\text{scan}}(\mathbf{v}_i) = \frac{1}{\sigma_1} \sum_{k=1}^M |I_{i,k} - F(\mathbf{p}_{i,k}, \mathbf{v}_{i-1}, \mathbf{v}_i, I_{i-1})|. \quad (5)$$

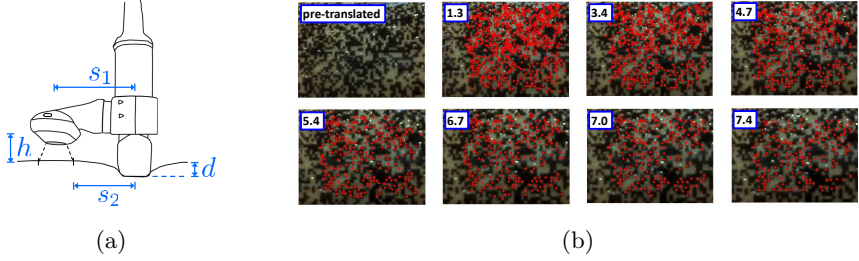
*Energy Minimization.* The global optimal estimate  $\mathbf{v}_i^*$  minimizes the total energy  $E_{\text{total}}(\mathbf{v}_i)$ , where  $E_{\text{total}} = E_{\text{prior}} + E_{\text{reproj}} + E_{\text{scan}}$ . The variances were determined empirically.  $\mathbf{v}_i^*$  was found by using the Levenberg-Marquardt algorithm.

## 2.4 Robustness to Probe Compression

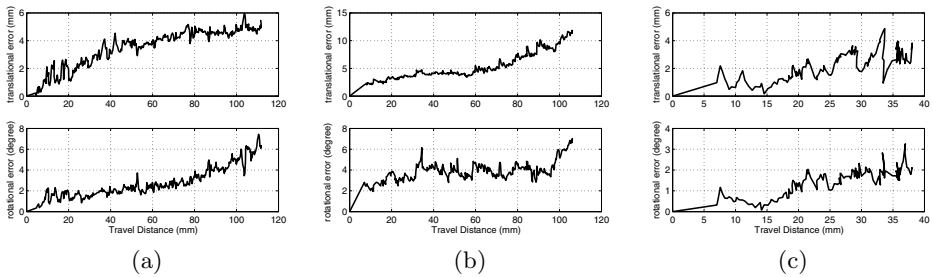
Visual SLAM could suffer from probe contact, which locally deforms the skin surface. However, with a careful hardware design, this deformation could be avoided in the camera FOV and hence visual SLAM could essentially be applied on the undeformed surface. As shown in Fig. 3(a), the distance between the FOV and probe contact ( $s_2$ ) could prevent deformation from being visible in the FOV. Although increased compression causes more deformation, it also brings the FOV farther away from probe contact. As a result, local surface deformation could be mostly avoided in the FOV even under significant probe compression.

Our design ( $h = 27$  mm,  $s_1 = 42.6$  mm, and  $s_2 = 32.1$  mm with no probe compression) was found to work well for avoiding surface deformation in the camera FOV, which was validated by experiments performed on the abdomen of a human subject. A temporary tattoo was affixed to the abdomen to ensure rich and high-contrast features all over the camera FOV. A camera image was first taken with minimal probe compression on the abdomen. The probe was then translated by 4.5 mm to a new position, where camera images were taken under varying compression. Between the pre-translated image and each one of the post-translated images, SIFT feature correspondences were found and the five-point algorithm was performed within a RANSAC scheme, which is similar to visual SLAM initialization (Section 2.3). Note that since the five-point algorithm determines the relative pose based on a rigid motion model, skin features that underwent local non-rigid deformation would be rejected as outliers in RANSAC.

The pre- and post-translated camera images are shown in Fig. 3(b) along with the inlier feature points. As can be seen, for all the compression levels, inlier points are present all over the FOV. This observation suggests that the skin features in the FOVs underwent little non-rigid deformation during probe compression, including those features close to probe contact (near the bottom of images). It is therefore confirmed that local surface deformation is avoided in the FOV by this hardware design even when the surface is highly compressed.



**Fig. 3.** (a) Illustration of skin surface deformation due to probe contact. (b) The pre-translated camera image and post-translated images corresponding to varying compression distances against the skin surface (in mm; denoted by  $d$  in (a)). Note that for all the compression levels, the inlier feature points (marked in red) are present all over the FOV except the leftmost portion, which is not covered by the pre-translated FOV.

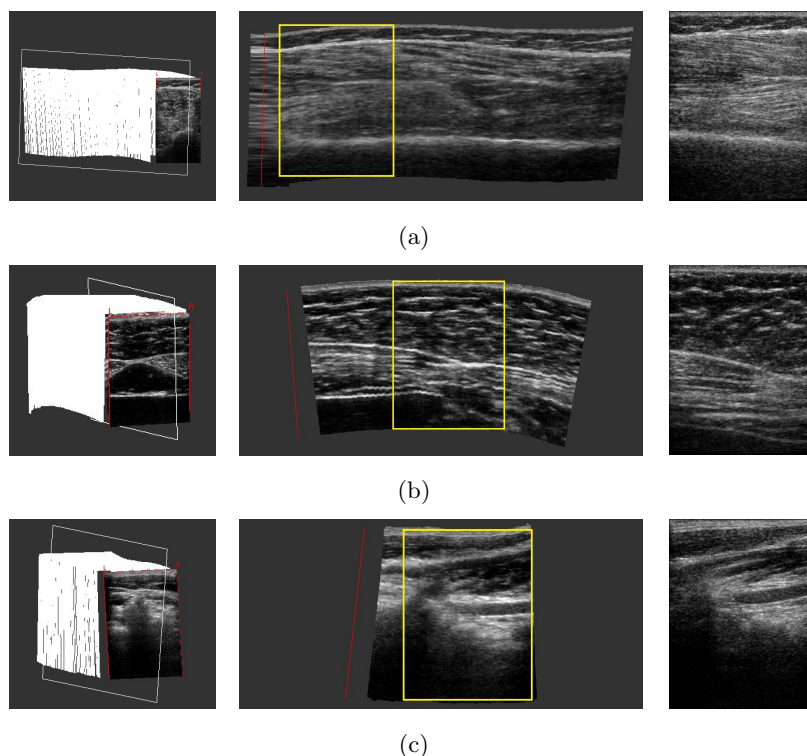


**Fig. 4.** Motion errors in translation (top) and rotation (bottom) for the freehand scan on (a) lower leg, (b) abdomen, and (c) neck of a human subject

### 3 *In-Vivo* Experiments

Our system was validated by freehand scanning on three body parts of a human subject, including the lower leg, abdomen, and neck. As the ground truth, independent measurement of probe motion during these scans was obtained using OptiTrack V120:Trio (NaturalPoint Inc.), a tri-camera 6-DoF tracking device with sub-millimeter accuracy. The tracked probe poses from our system and V120:Trio were then compared in the world coordinate system of V120:Trio. The body parts were immobilized during scanning to allow this comparison.

Five freehand scans were performed on each body part and examples of the pose estimation errors from our system are shown in Fig. 4. The drifting effect is observable, which is common for dead-reckoning approaches and is a short-coming of our system compared to optical or EM tracking. On average, for every probe travel distance of 10 mm, the accumulated error is  $0.91 \pm 0.49$  mm in translation and  $0.55^\circ \pm 0.17^\circ$  in rotation. A major cause of the estimation errors is the well-known ambiguity in discriminating between translation and rotation from camera frames [1]. Other sources of errors include inaccuracies in spatial



**Fig. 5.** Volume reconstructions and validation for the freehand scan on (a) lower leg, (b) abdomen, and (c) neck of a human subject. Left: reconstructed 3D volumes and reslice planes, with the transducer at the top. Middle: synthesized reslices. Right: real US scans acquired at approximately the same positions and orientations as the reslice planes.

calibration and body motion during scanning, such as motion caused by tremor, respiration, and heartbeat, especially for scans on the abdomen and neck.

By using the localization results from our system, 3D US volumes were reconstructed and visualized using Stradwin [13]. Examples of these 3D reconstructions and the validation are shown in Fig. 5, where probe pressure and motion jitter were corrected in Stradwin. Note that the reslice from the abdomen scan appears to be more noisy, which is mainly due to physiological tissue motion such as respiration and heartbeat. These synthesized reslices were validated by real US scans acquired at approximately the same positions and orientations as the reslice planes, which are shown on the right of Fig. 5. It can be seen that the tissue structures are consistent between the real US scans and the portion of reslices highlighted by yellow. This consistency shows the potential practical value of our system in improving clinical workflows and aiding diagnosis by enabling the creation of diagnostically useful reslices after US scanning.

## 4 Conclusion

We have presented a highly cost-effective and miniature-mobile system for free-hand 3D US, which localizes the probe by tracking skin features and is therefore robust to rigid patient motion. Results from freehand scans on three body parts show that the system could potentially be an alternative to conventional tracking devices for freehand 3D US with a much lower cost and higher portability.

**Acknowledgement.** The authors thank Singapore-MIT Alliance, GE Global Research, and Terason for the support on this work.

## References

1. Adiv, G.: Inherent Ambiguities in Recovering 3-D Motion and Structure from a Noisy Flow Field. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 11(5), 477–489 (1989)
2. Flaccavento, G., Lawrence, P., Rohling, R.: Patient and Probe Tracking during Freehand Ultrasound. In: Barillot, C., Haynor, D.R., Hellier, P. (eds.) *MICCAI 2004. LNCS*, vol. 3217, pp. 585–593. Springer, Heidelberg (2004)
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), 137–149 (2006)
4. Goldsmith, A.M., Pedersen, P.C., Szabo, T.L.: An Inertial-Optical Tracking System for Portable, Quantitative, 3D Ultrasound. In: *IEEE International Ultrasonics Symposium (IUS)*, pp. 45–49 (2008)
5. 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. In: Linte, C.A., Moore, J.T., Chen, E.C.S., Holmes III, D.R. (eds.) *AE-CAI 2011. LNCS*, vol. 7264, pp. 58–64. Springer, Heidelberg (2012)
6. Lepetit, V., Moreno-Noguer, F., Fua, P.: EPnP: An Accurate  $O(n)$  Solution to the PnP Problem. *International Journal of Computer Vision* 81(2), 155–166 (2009)
7. Lowe, D.: Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision* 60(2), 91–110 (2004)
8. Nistér, D.: An Efficient Solution to the Five-Point Relative Pose Problem. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 26(6), 756–770 (2004)
9. Prager, R., Rohling, R., Gee, A., Berman, L.: Rapid Calibration for 3-D Freehand Ultrasound. *Ultrasound in Medicine & Biology* 24(6), 855–869 (1998)
10. Rafii-Tari, H., Abolmaesumi, P., Rohling, R.: Panorama Ultrasound for Guiding Epidural Anesthesia: A Feasibility Study. In: Taylor, R.H., Yang, G.-Z. (eds.) *IP-CAI 2011. LNCS*, vol. 6689, pp. 179–189. Springer, Heidelberg (2011)
11. Stolka, P.J., Kang, H.J., Choti, M., Bector, E.M.: Multi-DoF Probe Trajectory Reconstruction with Local Sensors for 2D-to-3D Ultrasound. In: *IEEE International Symposium on Biomedical Imaging (ISBI)*, pp. 316–319 (2010)
12. Sun, S.-Y., Gilbertson, M., Anthony, B.W.: 6-DOF Probe Tracking via Skin Mapping for Freehand 3D Ultrasound. In: *IEEE International Symposium on Biomedical Imaging (ISBI)*, pp. 780–783 (2013)
13. Treece, G.M., Prager, R.W., Gee, A.H.: The Stradwin 3D Ultrasound Acquisition and Visualisation System, <http://mi.eng.cam.ac.uk/~rwp/stradwin/>