Ultrasound Probe Localization
by Tracking Skin Features
by
Shih-Yu Sun
B.S., National Taiwan University (2006)
S.M., Massachusetts Institute of Technology (2010)
Submitted to the
Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
at the
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
September 2014
© Massachusetts Institute of Technology 2014. All rights reserved.
Ultrasound Probe Localization by Tracking Skin Features

by

Shih-Yu Sun

Submitted to the Department of Electrical Engineering and Computer Science on August 22, 2014, in partial fulfillment of the requirements for the degree of Doctor of Philosophy

Abstract

Ultrasound probe localization with respect to the human body is essential for free-hand three-dimensional ultrasound (3D US), image-guided surgery, and longitudinal studies. Existing methods for probe localization, however, typically involve bulky and expensive equipment, and suffer from patient motion artifacts.

This thesis presents a highly cost-effective and miniature-mobile system for ultrasound probe localization in six degrees of freedom that is robust to rigid patient motion. In this system, along with each acquisition of an ultrasound image, skin features in the scan region are recorded by a lightweight camera rigidly mounted to the probe. Through visual simultaneous localization and mapping (visual SLAM), a skin map is built based on skin features and the probe poses are estimated. Each pose estimate is refined in a Bayesian probabilistic framework that incorporates visual SLAM, ultrasound images, and a prior motion model.

Extraction of human skin features and their distinctiveness in the context of probe relocalization were extensively evaluated. The system performance for free-hand 3D US was validated on three body parts: lower leg, abdomen, and neck. The motion errors were quantified, and the volume reconstructions were validated through comparison with ultrasound images. The reconstructed tissue structures were shown to be consistent with observations in ultrasound imaging, which suggests the system’s potential in improving clinical workflows.

In conjunction with this localization system, an intuitive interface was developed to provide real-time visual guidance for ultrasound probe realignment, which allows repeatable image acquisition in localized therapies and longitudinal studies. Through in-vivo experiments, it was shown that this system significantly improves spatial consistency of tissue structures in repeated ultrasound scans.

Thesis Supervisor: Brian W. Anthony
Title: Principal Research Scientist, Department of Mechanical Engineering

Thesis Supervisor: Charles G. Sodini
Title: Clarence J. LeBel Professor of Electrical Engineering
Acknowledgments

Graduate studies at MIT would not have been such an enjoyable and exciting journey without the support of my advisors, colleagues, collaborators, friends, and family. Specifically, I am immensely grateful to my advisor, Dr. Brian W. Anthony, for his guidance during the past six years. He has been a wonderful mentor, colleague and friend, and he has established an amazing environment for me to learn, explore, work and grow. This thesis work would not be possible without his support.

I am incredibly fortunate to have Professor Charles G. Sodini as my co-advisor, and Professor Berthold K.P. Horn and Professor George C. Verghese as my thesis committee members. They are extremely experienced researchers, engineers and educators. Their insights and expertise have inspired me to look at the research work from different perspectives, which gave me opportunities to explore various possibilities in development and analysis of the work.

Dr. Seward Rutkove from Beth Israel Deaconess Medical Center and Dr. Anthony Samir from Massachusetts General Hospital have been great clinical collaborators. They have provided great insights and suggestions on my thesis work as both professional practitioners and users. Working with them has allowed me to improve usability of the prototypes in clinical settings.

This thesis work has been financially and technically supported by GE Global Research, Singapore-MIT Alliance (SMA), and Terason. Discussions with my collaborators from GE Global Research, including Kai E. Thomenius, Aaron M. Dentinger, and Jeffrey M. Ashe, have always been helpful and inspiring.

The support from my lab members and the fun we had have also been an invaluable part to my life at MIT. I feel really lucky to have had the chance to work alongside with Matthew, Dean, Nadege, Aaron, Sisir, John, Xian, Ina, Shawn, and Nigel. With the diverse background, they have taught me a variety of things that I could not possibly have learnt by myself and shown me things that I could not imagine.
Finally, I would like to thank my family in Taiwan, especially my mother, Shu-Huei Chen, and my father, Kao-Chin Sun. Their love and support have been extremely important to me. I am also grateful to my wife, Yi-Ping Jian, for everything she has done for me. I have learnt a lot from her, and she has made my life at MIT much more enjoyable. I deeply appreciate it and will never forget.
Contents

1 Introduction 21
  1.1 Three-Dimensional Ultrasound 22
  1.2 Spatial Registration of Ultrasound Images 23
    1.2.1 Mechanically Actuated Probe 23
    1.2.2 Two-Dimensional Transducer Array 24
    1.2.3 Tracked Freehand Scanning 25
  1.3 Related Work on Tracked Freehand Scanning 27
  1.4 Related Work on Ultrasound Probe Realignment 30
  1.5 Proposed System for Probe Localization 32
  1.6 Contributions 33
  1.7 Summary 34

2 Probe Localization and Relocalization 37
  2.1 Planar Skin Surface Mapping: Homographies 39
    2.1.1 World-to-Image Planar Homographies 39
    2.1.2 Computing Homography Matrices 40
    2.1.3 Recovering Camera Rotation and Translation 41
    2.1.4 Experimental Results 42
  2.2 General Skin Surface Mapping: Visual SLAM 43
    2.2.1 Two-Frame Initialization 45
    2.2.2 Camera Pose Tracking 47
    2.2.3 Map Extension 47
    2.2.4 More on RANSAC 48
## In-Vivo 3D Ultrasound Experiments

6.1 Nearly Planar Surface with Natural Skin Features .................................. 93
6.2 General Skin Surface with Artificial Skin Features ................................. 95
6.3 General Skin Surface with Natural Skin Features .................................. 98
   6.3.1 Experimental Setup ............................................................... 98
   6.3.2 Motion Errors ........................................................................ 101
   6.3.3 Skin Feature Retrieval for Probe Relocalization ......................... 103
   6.3.4 3D Volume Reconstructions and Reslices .................................. 105
6.4 Summary ......................................................................................... 106

## Computer-Guided Ultrasound Probe Realignment

7.1 Two-Frame Initialization ........................................................................ 110
7.2 Camera Pose Tracking ......................................................................... 111
7.3 Visual Guidance .................................................................................. 112
7.4 Implementation .................................................................................... 113
7.5 In-Vivo Experimental Results ............................................................. 114
7.6 Summary ............................................................................................. 117

## Conclusion

8.1 Contributions ....................................................................................... 119
8.2 Limitations and Future Work ............................................................... 121
   8.2.1 Defocus Blur ............................................................................. 121
   8.2.2 Fusion of Multiple Information Sources ................................. 122
   8.2.3 Discrete-Continuous Optimization ........................................ 125
   8.2.4 Probe Realignment based on Natural Skin Features .............. 128
List of Figures

1-1 Illustration of the procedure for tracked freehand scanning 3D US (figures from [18, 19, 30]). ................................. 26

1-2 In the proposed probe localization system, the skin features are recorded with each ultrasound image acquisition. .............. 32

2-1 Summary of the scan process and the two algorithms for simultaneous camera tracking and skin mapping. .................. 38

2-2 An experiment was performed to validate the tracking and mapping algorithm based on planar homographies. In the setup (a), the camera moved along the grid and captured a portion of the pattern at each stop. In the results (b), it can be seen that the camera position estimates (blue) are very close to the ground truth (red). The translational error was about 1 mm on average. .................. 43

2-3 Flowchart for the visual SLAM reconstruction process. ............ 44

2-4 (a) Probability of success in RANSAC against the inlier ratio given varying numbers of samples. (b) Required number of samples in RANSAC against the inlier ratio given desired probabilities of success. .................. 49

2-5 A square sticker with known dimensions that is affixed to the skin surface of the scan region for scale calibration. .................. 51

2-6 A pattern with random dots and a calibration square on the left. . 52
Experiments were performed to validate the visual SLAM algorithm. In the planar case (a), the camera moved along a straight line. In the cylindrical case (b), the camera moved along a straight line initially by 1 cm. Then the camera was fixed, and the cylindrical object rotated to create relative motion.

The experimental results of the visual SLAM algorithm on (a) planar and (b) cylindrical objects, showing the scene map on the left and camera trajectory on the right. It can be seen that both results are consistent with the experimental setups.

The portion of an ultrasound image (shown in gray) used for computation of image regularity, which is indicated by the blue region.

(a) A random binary pattern as artificial skin features. A square with known dimensions is added to this pattern for scale calibration. (b) A human subject’s thigh covered by the transparent dressing and then the temporary tattoo.

Natural skin features covered by evenly applied ultrasound 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. 4-1.

The performance of matching contrast-enhanced skin features between two camera views was examined on the forearms of four human subjects with varying camera exposure levels and thus varying image SNRs. (a) Skin tones of the human subjects at the forearms under the same lighting condition with exposure level 3. (b) The same forearm region of subject A under the same lighting condition with four different exposure levels. Level 1 gave the shortest exposure time and level 4 the longest. (The images shown here are not contrast-enhanced.)
3-4 Performance of matching contrast-enhanced skin features between two camera views on four human subjects with varying camera exposure levels, as shown in Fig. 3-3. (a) Number of all SIFT feature correspondences. The error bars indicate standard deviations over ten frame pairs. (b) Number of inlier feature correspondences. (c) Inlier ratios.

3-5 Different color channels (i.e. RGB, YIQ, and HSV) from the skin images of the four human subjects shown in Fig. 3-3(a).

3-6 The performance of matching contrast-enhanced skin features between two camera views was examined on the forearms of four human subjects by using different channels in three color representations: RGB, YIQ, and HSV. The subjects were the same as shown in Fig. 3-3 and the camera exposure was fixed at level 3. (a) Number of all SIFT feature correspondences. The error bars indicate standard deviations over ten frame pairs. (b) Number of inlier feature correspondences. (c) Inlier ratios.

4-1 The ultrasound probe with a rigidly mounted camera for recording skin features.

4-2 The checker-board pattern for camera calibration. The square size is 12×12 pixel, which amounts to around 0.5×0.5 mm when printed with 600 dots per inch (DPI). The white dot at the top left marks the origin of the coordinate system in calibration.

4-3 (a) Illustration of the “2D-target method” for ultrasound calibration. (b) The phantom designed to perform the 2D-target method.

4-4 (a) The metal structure embedded in the calibration phantom for the 2D-target method. (b) A camera calibration pattern for estimating the camera-to-world transformation.
4-5  (a) Flat-bottom water tank with a camera calibration target for single-wall ultrasound calibration. (b) One of the ultrasound images for calibration. The detected line representing the flat bottom is shown in green.

4-6  The ultrasound probe with a housing for the camera and lighting source.

4-7  Skin features recorded by the camera shown in Fig. 4-6 before (top) and after (bottom) contrast enhancement, as described in Section 3.2: (a) no light, (b) weak light, and (c) strong light. (The camera exposure time in (a) is four times that in (b) and (c).)

4-8  Illustration of specular reflection from the skin surface covered by ultrasound gel. Note that when the incident angle of light is small with respect to the skin surface normal, the reflected light could enter the camera aperture, which results in specular reflection.

4-9  The same skin surface seen in (a) visible light and (b) visible+IR light. The original images are shown at the top and the contrast-enhanced images (as described in Section 3.2) are shown at the bottom. Note that vessels in (b) are slightly more obvious.

4-10 Illustration of skin surface deformation due to probe compression. Distance measurements before (left) and after (right) probe compression are indicated. Note that the distance between the camera FOV and probe contact increases with probe compression (hence $s_2' > s_2$).

4-11 The pre-translated camera image and post-translated images corresponding to varying compression distances (in mm) against the skin surface of the abdomen of a human subject. Black dots shown in the images are artificial skin features created in the scan region. Note that for all the compression levels, 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.
5-1 (a) Artificial surface texture on the phantom for in-vitro experiments on a planar surface. The square in (b) is also shown. (b) A square sticker with known dimensions to define the world coordinate system. The four corners (marked by red circles) are located with sub-pixel accuracy in the camera images. ........................................ 88

5-2 (a) The phantom dimensions in the in-vitro experiments based on planar homographies. (b) An example ultrasound image from linear scanning on the phantom. ................................. 89

5-3 (a) Illustration of the three scanning types: linear, tilt, and rotational scanning. (b)-(d) Examples of the scan trajectories and reconstructed cylinders. ............................. 90

5-4 Comparison of mean volume errors. ........................................... 90

5-5 (a) The phantoms with planar (top) and curved (bottom) surfaces, covered by temporary tattoo stickers. (b) Estimated surface feature points and camera trajectory (shown in blue) from scanning the phantom in (a). Part of the corresponding ultrasound image trajectory is shown at the bottom right (in gray). (c) Ultrasound scan trajectory and the reconstructed cylinder, which is highlighted in the inset. ......................................................... 91

6-1 In-vivo experimental setup on a nearly planar skin surface. .......... 94

6-2 In-vivo experimental results based on planar homographies. (a) Aligned skin features and the tenth camera pose as an example. (b) Acquired ultrasound images registered in the world coordinate system. ......................................................... 94

6-3 3D reconstruction of the brachial artery. (a) One of the ultrasound images showing the segmented artery. (b) Volume reconstruction of the brachial artery. The ultrasound scan trajectory is also shown. 95
6-4 (a) A human subject’s forearm covered with the temporary tattoo was scanned. (b) A camera video and an ultrasound video were recorded synchronously during scanning.

6-5 (a) Reconstructed skin surface and camera poses. The blue cameras correspond to keyframes, and the white camera is the one currently tracked. (b) The ultrasound images spatially registered in 6 DoF based on the camera pose estimates. (unit: mm)

6-6 (a) One of the ultrasound images showing the segmented artery. (b) Volume reconstruction of the brachial artery. The ultrasound scan trajectory is also shown.

6-7 (a) The scan region with artificial skin features in the in-vivo experiments. (b) Estimated skin feature points and camera trajectory (shown in blue) from scanning the region in (a). The corresponding ultrasound scan trajectory is also shown (in gray). (c) One of the recorded ultrasound images. The femoral artery (enclosed by the yellow rectangle) was manually segmented and surface-rendered. The reconstructed artery is shown in the inset.

6-8 Body parts where freehand scanning experiments were performed and the scan paths (shown in blue): (a) lower leg, (b) abdomen, and (c) neck.

6-9 (a) Setup for linear scanning experiments on the lower leg of a human subject. Probe motion was independently measured by OptiTrack V120:Trio as the ground truth. Involved coordinate transformations (green) and the coordinate system $U_1$ (black) are illustrated. (b) Trajectories of the ultrasound image origin in $U_1$ (in mm): ground truth (black), prior+camera+ultrasound (blue), and prior+camera (red).

6-10 Motion errors in translation (top) and rotation (bottom) in the freehand scan on (a) lower leg, (b) abdomen, and (c) neck of a human subject.
6-11 Estimation errors in probe motion versus the total probe travel distance (in mm) in the lower-leg experiment: prior+camera+ultrasound (blue) and prior+camera (red, dashed). (a) Translational errors along the X, Y, and Z axes (in mm) in coordinate system U1. (b) Rotational errors around the three axes (in degrees).

6-12 The (a) precision and (b) recall for skin feature retrieval in an example scan. Note the spikes in recall, which happen at keyframes.

6-13 Example reslice from the lower leg scan (a) before and (b) after correction of probe pressure and jitter in motion estimation.

6-14 Visualization and validation of the freehand scan on (a) lower leg, (b) abdomen, and (c) neck of a human subject. Left: reconstructed 3D volumes and reslice planes. Middle: synthesized reslices. The red lines correspond to the ultrasound images acquired at time instance $i = 1$. Right: real ultrasound images acquired at approximately the same positions and orientations as the reslice planes. Note that the tissue structures are consistent between the real ultrasound images and the portion of reslices highlighted by yellow.

7-1 (a) Illustration of the virtual pyramid augmented to the map of 3D feature points. (b) An example of the probe realignment process for the case in (a).

7-2 (a) An example binary random pattern as artificial skin features. (b) The scan region with artificial skin features.

7-3 The camera frames (top) and ultrasound images (bottom) acquired at the (a) starting, (b) realigned, and (c) target (ground truth) probe pose. The femoral artery is highlighted by yellow rectangles.
7-4  (a) The difference image between the starting and target ultrasound images (Fig. 7-3(a) and (c)). The regions where the target image is brighter are shown in green, and those darker are shown in red. (b) The difference image between the realigned and target ultrasound images (Fig. 7-3(b) and (c)). ........................................ 116

8-1  Two cameras mounted to the probe with non-overlapping FOVs for improvement of motion estimation. ................................. 123

8-2  Illustration of the MRF model. ......................................... 126

8-3  Calibration for ultrasound probe realignment based on natural skin features and a real-time tracker. ................................. 129
List of Tables

6.1 Summary of 3D Reconstructions for Freehand Scans ............... 104
6.2 Results of Skin Feature Retrieval and Probe Relocalization for Free-hand Scans ........................................... 104
Chapter 1

Introduction

Medical ultrasound imaging technology is indispensable nowadays as it provides safe, inexpensive, portable, and real-time imaging with high spatial resolution. Nevertheless, there are some limitations inherent in conventional two-dimensional (2D) imaging, including limited scanning angles, acquisition based on prior assumptions of tissue structures, and challenges in repeatable image acquisition in longitudinal studies. This thesis aims to address the limitations in 2D ultrasound imaging by designing a highly cost-effective, miniature-mobile system for ultrasound probe localization with respect to the patient's body.

In this chapter, the two main applications of the proposed system are introduced: three-dimensional ultrasound imaging (3D US) and computer-guided ultrasound probe realignment. The advantages of 3D US over 2D US are first discussed in Section 1.1. The major approaches for performing 3D US are then reviewed and categorized in Section 1.2. Specifically, the proposed system falls into the category of “tracked freehand scanning”, which is discussed in greater detail in Section 1.3. Finally, the related work on ultrasound probe realignment is reviewed in Section 1.4. After these discussions on the state of the art, the proposed system is described and the contributions are emphasized.
1.1 Three-Dimensional Ultrasound

In the 3D ultrasound scanning procedure, 2D ultrasound images are densely acquired and accurately registered in space to reconstruct a 3D volume of the underlying tissue. This volume reconstruction enables volume rendering or surface rendering of the tissue structures, and therefore is especially useful for visualizing complex structures, which often happen in gynaecological, neck, anorectal, skin, kidney, and brain imaging [91, 27, 28, 76]. Acquiring the whole 3D volume in ultrasound imaging also allows easier registration with images from other imaging modalities, such as computed tomography (CT) and magnetic resonance imaging (MRI) [53].

Another advantage of 3D US is the reduction of operator dependence in ultrasound image acquisition and flexibility in extracting anatomical information at the diagnosis stage. In 2D imaging, the location and orientation of image acquisition depends on the operator’s interpretation of the underlying tissue structures, which involves strong prior assumptions. In contrast, 3D US allows separation of image acquisition and interpretation. Given a whole 3D volume of the tissue, 2D ultrasound images that are needed for diagnosis, and even those that are from angles not possible in conventional imaging, could later be synthesized simply by re-slicing the volume. As a result of this consistency in acquisition, 3D US provides more accurate and repeatable quantitative measurement of tissue and is therefore more suitable for monitoring the progression of diseases [15, 3, 26].

In addition to diagnostic visualization, 3D US is also useful in supporting surgical interventions [85]. For instance, 3D US provides guidance in prostate biopsy and brachytherapy, where biopsy needles or radioactive seeds can be placed interactively in the 3D volume [12, 96, 122]. Also, in radio-frequency (RF) ablation of hepatic tumors, 3D US allows spatial registration of intraoperative ultrasound images, which visualize the ablation target area in real time, with preoperative CT images that are used for delineating tissue anomalies [53]. [91] provides an extensive overview of the clinical applications improved or enabled by 3D US.
Currently, 3D US has been incorporated into some high-end ultrasound imaging equipment to provide advanced imaging capabilities. However, the adoption of 3D US in hospitals has been limited, partially because the equipment is expensive or difficult to use as discussed in Section 1.2. Aiming to make 3D US technology more accessible, this thesis presents the design and analysis of a cost-effective and mobile system for six-degrees-of-freedom (6-DoF) spatial registration of ultrasound images by tracking patient skin features with a camera rigidly mounted to the probe. It is envisioned that the proposed system could facilitate the use of advanced medical ultrasound technologies for improved diagnostic capabilities and remote diagnosis, especially in underdeveloped regions and countries.

1.2 Spatial Registration of Ultrasound Images

Accurate spatial registration of acquired 2D ultrasound images is one of the essential components of 3D US, which generally requires determining 6-DoF spatial rigid transformations of the images with respect to each other. There are currently three major approaches with which 2D ultrasound images are spatially registered: mechanically actuated probes, 2D transducer arrays, and tracked freehand scanning. In this section, these approaches are briefly introduced and discussed.

1.2.1 Mechanically Actuated Probe

Currently the most common type of commercial 3D US systems is that based on mechanically actuated probes. A stepper motor is compactly integrated inside the probe housing to drive a 1D transducer array similar to that used in conventional imaging. The transducer is acoustically coupled with the patient's skin surface through a bag of oil inside the housing. During scanning, the operator fixes the probe on the patient's skin surface and 2D images are then acquired following a pre-determined trajectory, making the spatial transformations between images readily available. See [91] for an example of the mechanical design, where the
Kretztechnik RSP6-12 probe manufactured by GE Medical Systems is described in detail.

Mechanically actuated probes provide real-time acquisition of volumes with an image quality comparable to that in conventional imaging since they use similar 1D transducer arrays. This use of conventional transducers also makes the probes cheaper to support than the two-dimensional transducer arrays described in the following section. As a result, mechanically actuated probes are currently the most popular approach for 3D US, especially in obstetrics and gynaecology (OB/GYN) applications.

1.2.2 Two-Dimensional Transducer Array

2D piezoelectric transducer arrays were first introduced in the 1990s [99]. As opposed to conventional 1D transducer arrays, which acquire a single 2D image at a time, 2D transducer arrays allow parallel acquisition of spatially registered images. During scanning, the operator places the 2D array probe on the patient body, and a pyramid-shape volume is acquired. As ultrasound beams are steered and focused electronically, this approach enables very accurate spatial registration of images. With computational capabilities for parallel beamforming, this approach also allows high-speed volume acquisition and therefore is particularly useful in cardiac imaging.

In recent years, fabrication of 2D transducer arrays and efficient signal processing of the massive data are under active development. For instance, [126] introduces a 5-MHz 128×128 rectilinear array, in which each element is 600 \( \mu \text{m} \times 600 \mu \text{m} \). [55] describes a 3-MHz 48×48 sparse transducer array, where the number of elements is reduced compared to the full array while the aperture is maintained. Commercial 2D arrays are also being manufactured. For instance, Philips now offers 2D transducer arrays with central frequencies ranging from 1 to 7 MHz with up to 3040 elements (around 55×55) to support their high-end imaging systems.

Despite the development of 2D transducer array technologies, a number of
challenges in manufacturing such arrays need to be addressed before they could become widely available. For instance, it is difficult to individually wire and interconnect the piezoelectric elements without distorting the mechanical structure of the transducer assembly. Also, since the elements in 2D arrays are smaller than those in conventional 1D transducers, impedance matching between the elements and the coaxial cable becomes more difficult, and the image quality tends to be poorer [108]. Furthermore, the need to improve impedance matching and the signal-to-noise ratio (SNR) results in a massive amount of electronics in the front end and makes the transducer head bulkier. In order to address these issues, micromachined ultrasonic transducers (MUT) are under active development, including capacitive transducers (CMUT) and piezoelectric transducers (PMUT), which provide a wider bandwidth and allow easier fabrication. More details on the development of MUT could be found in [83, 4, 84].

1.2.3 Tracked Freehand Scanning

Both the mechanically actuated 1D arrays and 2D transducer arrays provide real-time volume acquisition and are easy to use. However, it is expensive to apply these approaches as dedicated probes with special hardware design are required. Also, the scanned region is physically constrained by the form factor of the probe since the probe needs to be fixed during scanning. Without additional motion tracking facilities, it is difficult to perform acquisition of large volumes, such as in scanning the liver or the whole abdomen.

In contrast to the approaches discussed above, 3D US based on tracked freehand scanning allows volume acquisition of an arbitrary scan geometry with a higher voxel resolution. This approach is also more readily accessible since it could be performed directly by using existing 2D ultrasound imaging systems. Hence, the use of freehand 3D ultrasound has been reported in a wide range of clinical applications [52, 91, 29]. Specifically, freehand 3D US is particularly useful for vascular and musculoskeletal imaging as it allows the use of existing
Tracked freehand Spatial registration of 3D volume ultrasound scanning ultrasound Images reconstruction

Figure 1-1: Illustration of the procedure for tracked freehand scanning 3D US (figures from [18, 19, 30]).

high-frequency ultrasound transducers. In freehand 3D US, the operator moves a conventional ultrasound probe in freehand over the anatomy as in a conventional procedure while 6-DoF motion of the probe is being tracked. The position and orientation of the acquired 2D images are then estimated for volume reconstruction of the underlying tissue structures, as illustrated in Fig. 1-1.

Currently in clinical practice, it is common to perform freehand scans without tracking probe motion and then assemble the 2D ultrasound images into a 3D volume for subsequent diagnosis. This approach relies on the assumption that the probe motion follows a pre-determined scan trajectory (usually a line) and the scan speed remains constant throughout the process. Therefore in this practice, probe motion is highly constrained. Additionally, since it is difficult to fully meet these assumptions by freehand scanning, the 3D volumes acquired in this manner are also unreliable for diagnostic interpretation based on quantitative measurement of tissue dimensions.

For accurate reconstruction of 3D tissue volumes, localization of the ultrasound probe with respect to the patient's body in 6 DoF is essential. The probe localization system presented in this thesis falls into this category, and the related
work is discussed in greater detail in the following section.

1.3 Related Work on Tracked Freehand Scanning

Some earlier work in freehand 3D US used acoustic sensing for probe localization, which locates a number of transmitters attached to the probe in a fixed coordinate system by measuring the time of flight of sound waves [71]. There has also been development of tracked ultrasound scanning systems with movable mechanical tracking arms, where the probe is mounted to the multiple-jointed arm and moved in freehand. Motion of the probe is then estimated based on readings from rotary encoders installed at the arm joints [82].

Nowadays, one of the most popular methods for freehand probe localization involves the use of an optical tracker [69, 123, 116, 36, 20]. A passive or active target is rigidly attached to the ultrasound probe and is tracked by two or more cameras. As the operator performs freehand scanning, both the ultrasound images and probe motion are recorded synchronously for later spatial registration of the images. This approach is able to provide sub-millimeter accuracy in real time. A major drawback of using an optical tracker in freehand scanning is the need to maintain a direct line of sight between the optical system and the target, which limits the scanning trajectory. Additionally, the equipment required for optical tracking is typically bulky and expensive, which makes the use of this approach both economically and physically constraining.

Another popular method for probe localization involves the use of an external electromagnetic tracker [115, 14, 49, 13, 51, 75]. A transmitter of a time-varying electromagnetic field is placed near the patient. A sensor with three coils is rigidly attached to the ultrasound probe to pick up the electromagnetic signal as the probe moves, which allows estimation of the probe motion. Unlike using an optical tracker, the electromagnetic tracker does not require a line of sight between the transmitter and sensor. This approach is also able to provide sub-millimeter accuracy in real time. Nevertheless, ferromagnetic objects and electronic equipment
near the tracker could greatly affect the accuracy, which often makes application of this approach in clinical settings inconvenient [42]. The required equipment for this approach is also bulky and expensive as in optical tracking.

Patient motion artifacts in volume reconstructions are another concern when using an external tracking device with a world coordinate system that is decoupled with the patient’s body, such as in optical or electromagnetic tracking. In this case, the tracker localizes the probe in the world coordinate system and the anatomy under the freehand scan needs to remain still with respect to the tracker base throughout the scanning process for spatial consistency of the acquired 2D ultrasound images. When patient motion is not completely avoided during the freehand scan, spatial discontinuities are created in the reconstructed volume, which make diagnostic interpretation based on the acquisition unreliable.

There have been efforts to develop cost-effective methods for ultrasound probe localization based on optical sensing. The systems described in [61] and [5, 6] use a single camera to track high-contrast markers mounted to the probe or surgical instrument. Localization is then performed in the camera coordinate system, which is independent of the patient’s body and thus makes the system performance susceptible to patient motion. Probe localization with respect to the anatomy under a scan, and not just an independent coordinate system, could be achieved by tracking markers on both the probe and anatomy at the same time using a common external optical tracker, which then allows relating the probe motion to patient motion [32, 8]. However, in this approach, lines of sight need to be maintained between the tracker base and both tracking targets. Therefore, the use of this method is even more physically constraining than tracking the probe alone.

Instead of relying on external tracking devices, ultrasound probe localization by using only sensors mounted to the probe has been investigated, sometimes involving the aid of inertial measurement units (e.g. gyroscopes and accelerometers) for providing measurements in additional DoF or for refining estimates. These systems are more easily movable and usually more robust to rigid pa-
tient motion since localization could be performed directly with respect to the scan region. For instance, in [44], structured lighting sources and a camera are mounted to the ultrasound probe. By tracking the light pattern projected onto the skin surface during ultrasound image acquisition, the probe tilt angle against the skin surface is determined. [92] describes a probe tracking method that involves affixing a specialized strip with high-contrast markers to the patient’s body and moving the probe alongside the strip. Probe motion is estimated by tracking the markers using a camera mounted to the probe. Since the marker strip needs to be prepared beforehand, this approach is only suitable for simple and pre-determined scan paths. It is also possible to use two cameras mounted to the probe to track high-contrast patterns created on the entire surface of the scan region and then localize the probe with respect to the surface based on stereo vision, as demonstrated in the phantom studies in [121]. Nevertheless, these artificial patterns could affect ultrasound image quality and cause inconvenience in the clinical use of this method.

The systems described in [88, 39] and [101, 102] use probe-mounted optical systems similar to those in optical mice to track skin features from a small distance for determination of probe motion. As a result, only 2-DoF probe translation along the skin surface is estimated by these optical systems, although motion in other DoF could be measured by using inertial measurement units also mounted to the probe. These systems do not rely on markers and thus allow flexible scan paths. However, they were validated only on phantom materials and no results on human skin were reported. Also, the optical trackers must be in contact with the skin surface at all times during scanning, thus making application of this approach less convenient.

Ultrasound probe localization by using only ultrasound images is also under active research [60, 95, 21, 120, 46]. In this approach, relative probe motion between two adjacent images is estimated by measuring local correlation (or decorrelation) of ultrasound speckle patterns between the images. With the presence of fully developed speckles, a correlation-displacement curve for the specific tissue
under investigation could be measured, from which probe motion could be accurately estimated. Since only ultrasound images are used for the estimation, no spatial calibration is required between an additional positioning device and the ultrasound image plane. This approach, however, generally suffers from gradually increasing drifting error when it is iteratively applied to ultrasound image pairs along a sequence of images since the bias in each pose estimate accumulates. Attempts have been made to correct this drift by combining other sensing methods, such as optical tracking [47], electromagnetic tracking [59], and speckle tracking using additional ultrasound transducer arrays [45]. This method also relies on the existence of fully developed speckles to obtain 6-DoF motion estimates, so it tends to be inaccurate when applied on real tissue [37]. Overall, this approach is currently unable to provide tracking accuracy comparable to external motion sensors, so it is usually used only for qualitative, but not quantitative, volume imaging.

Finally, it is worth mentioning that the techniques for localization of an ultrasound probe are highly relevant to tracking endoscopes inside the patient's body. A popular method for tracking endoscopes is the electromagnetic tracker. Nevertheless, tracking endoscopes directly by using the organ surface texture in endoscopic images is also under active research and is highly relevant to the system presented in this thesis [127, 80, 66, 72, 114].

1.4 Related Work on Ultrasound Probe Realignment

In addition to diagnostic visualization, ultrasound imaging is also widely used in clinical settings to support surgical procedures. Specifically, ultrasound imaging is suitable for monitoring localized therapy, such as radio-frequency ablation, chemotherapy, and high-intensity focused ultrasound (HIFU), and for tracking progression of pathologies in longitudinal studies, partially due to its ease of use and real-time nature [85]. At each time point of examination in those processes, it is important to acquire ultrasound images at the same location, angle, and
compression level with respect to the body part, so that an unbiased comparison between images could be made for detecting tissue changes. However, accurate realignment of the ultrasound probe in 6 DoF with respect to the body remains challenging.

To obtain longitudinally repeatable ultrasound images of a particular region of interest, sonographers currently attempt to place the probe at the same body location and probe orientation with the aid of artificial fiducials on the skin surface, or by imaging at a measured distance from an anatomical feature. For instance, to monitor biceps muscle tissue, the sonographer might draw a line between the acromial process of the shoulder and lateral epicondyle of the elbow, and then image at the midpoint of the line, roughly perpendicularly to the skin surface. Although this method is convenient and inexpensive, the realignment accuracy is typically insufficient for detecting visually insignificant changes of tissue, especially when a slight deviation in probe orientation could result in substantial change of tissue appearance in the images.

Methods for automatic control of probe motion based on visual servoing are being developed to accurately track regions of interest or compensate tissue motion in ultrasound imaging, which are particularly useful in image-guided intervention. [1] describes a robotic probe control system that works alongside a human operator for carotid artery tracking. [58] and [68] describe systems that control probe motion purely based on ultrasound speckles and image moments, respectively, without human intervention. Promising results from these systems were reported, but the required equipment is expensive, bulky, and limited in portability.

Finally, re-positioning of a tool or system based on visual cues has been investigated beyond ultrasound imaging. This task is referred to as “visual homing” in robotics, which aims to design a system that automatically returns a robot to a desired position and orientation [10, 70]. Instead of automatic control, the system described in [9] provides visual guidance to help users realign a camera with respect to a previously captured photograph, which allows a comparison of the
In the proposed probe localization system, the skin features are recorded with each ultrasound image acquisition.

same scene over time, for instance.

1.5 Proposed System for Probe Localization

In this thesis, a highly cost-effective and miniature-mobile system is presented for ultrasound probe localization in the complete 6 DoF that is robust to rigid patient motion. A lightweight off-the-shelf camera is rigidly mounted to the probe to record natural or artificial skin features in the scan region at a small distance from the skin surface along with each acquisition of a 2D ultrasound image, as illustrated in Fig. 1-2. From the image sequence of skin features, a 3D skin map of the scan region is incrementally built and the 6-DoF camera pose corresponding to each camera frame is estimated with respect to the patient's body through visual SLAM (simultaneous localization and mapping). Finally, pose estimates of the camera are converted into those of the ultrasound images through the rigid spatial transformation between the camera and ultrasound image coordinates, which is found by spatial calibration.

In this system, additional sensors are not required for complete 6-DoF motion estimation although other sensors could be incorporated in a multi-estimator
framework. For instance, in order to obtain the optimal pose estimates, a Bayesian probabilistic framework has been developed in this thesis, which fuses information from visual SLAM, ultrasound images, and a prior motion model. Also, this method determines the probe poses with respect to the patient's body, and not an independent world coordinate system as in optical or electromagnetic tracking, so it is robust to rigid patient motion and allows spatial registration of separately acquired volumes.

One of the major applications of this probe localization system is freehand 3D US, and with the development of an intuitive user interface, computer-guided 6-DoF ultrasound probe realignment could also be achieved. The realignment system provides real-time visual guidance to sonographers for accurately returning the probe to a target pose at which a reference ultrasound image has been previously acquired. In this process, the system detects successful probe realignment and acquires ultrasound images automatically. This method could be thought of as an extension of the body fiducial approach to realign the probe in the complete 6 DoF with improved accuracy.

### 1.6 Contributions

The central contribution of this thesis is the development of a novel ultrasound probe localization system in 6 DoF. Compared to typical optical or electromagnetic tracking devices, this system is highly cost-effective and miniature-mobile since only an off-the-shelf camera is required. This system is also robust to rigid patient motion as probe poses are determined with respect to the skin surface, and not an independent world coordinate system. Two applications of this system are demonstrated and evaluated on human bodies: freehand 3D US and computer-guided ultrasound probe realignment.

A number of contributions were made in the development of the proposed system. Two novel ultrasound calibration procedures were designed, which are useful to systems involving both ultrasound imaging and camera sensing. A
Bayesian framework was developed to obtain the optimal pose estimates by integrating information from skin features, ultrasound images, and a prior motion model. Finally, a novel user interface was developed for ultrasound probe realignment, which involves creation of a static virtual pyramid amid the reconstructed skin map and rendering the projection in the camera frame as visual guidance for the sonographer.

Another major contribution of this work is the novel use of contrast-enhanced human skin features for probe localization with respect to the skin surface and probe relocalization using established skin map points as body landmarks. Through a series of experiments on human subjects, performance of skin feature identification from multiple camera views was evaluated for varying skin tones, body parts, and signal-to-noise ratios (SNR) of camera images. These results are important to the proposed system and potentially other applications utilizing natural body landmarks. Examples include tissue feature tracking in endoscopic images, where strong features are often absent and the lighting condition is highly dynamic [127]. The use of natural skin features also allows absolute spatial encoding of ultrasound images with respect to the patient’s body by recognition of distinctive skin features [106]. The system therefore enables repeatable image acquisition for disease monitoring, for instance.

1.7 Summary

Ultrasound probe localization in freehand scanning is essential for image-guided surgical intervention and three-dimensional imaging. Current approaches for probe localization, however, generally require bulky and expensive equipment, are sensitive to patient motion, and could be inconvenient to apply in clinical settings. This thesis presents a highly cost-effective and miniature-mobile probe localization system that is robust to rigid patient motion, and two applications of this system are demonstrated: freehand 3D US and computer-guided ultrasound probe realignment. It is envisioned that the proposed system could facilitate the
development of freehand 3D US and quantitative longitudinal studies.

This thesis is structured as follows. Chapter 2 gives an overview of the algorithms for mapping the skin surface and estimating camera motion at the same time by using skin features. The use of artificial and natural skin features for this task is discussed in Chapter 3. Design of the hardware, including the camera and probe housing, is discussed in Chapter 4, as well as related design considerations, including lighting and skin surface deformation due to probe contact. Application of the probe localization system to freehand 3D US is demonstrated both in vitro and in vivo in Chapters 5 and 6, respectively. Application of the system to ultrasound probe realignment is presented in Chapter 7. Finally, Chapter 8 concludes this work and the future work is discussed in this chapter.
Chapter 2

Probe Localization and Relocalization

The task of simultaneous camera tracking and skin mapping is often referred to as structure from motion or visual simultaneous localization and mapping (SLAM). Some examples of these algorithms are monoSLAM [25], PTAM [56], Bundler [100], and DTAM [77]. In this thesis, two such algorithms based on sparse image features are evaluated. Section 2.1 presents an algorithm based on planar homographies, which is suitable for planar or nearly planar skin surfaces [98]. The other algorithm for more general skin mapping is described in Section 2.2 and could be applied to skin surfaces of any shape. In these descriptions, it is assumed that the keypoints and feature descriptors have been extracted from the skin features by using the SIFT method [65], which could also establish the initial keypoint correspondences between images when necessary. The two algorithms are detailed in the following sections and summarized in Fig. 2-1.

Extensions of these algorithms are also presented in this chapter. For instance, the camera pose estimates obtained from these algorithms could be refined by combining information from multiple sources. Section 2.3 presents a Bayesian framework that incorporates skin features, ultrasound images, and a prior motion model for optimal estimation. Also, Section 2.4 describes an efficient algorithm for probe relocalization with respect to the previously established map, which is useful for recovering loss of tracking and for longitudinal studies.

Part of this chapter has been published in [104, 105, 107].
**Scan Process**

1. Perform intrinsic camera calibration (Section 4.1) and ultrasound calibration (Section 4.2) if needed.

2. Affix artificial skin features or a square with known dimensions (if natural skin features are used) to the skin surface of the scan region.

3. During scanning, skin features and ultrasound images are synchronously recorded.

**Off-Line Reconstruction**

1. Perform camera tracking and skin mapping for planar or general skin surfaces:
   - **planar surface**: (Section 2.1)
     i) In a camera image (denoted by \( p \)) where the square is fully visible, mark the four corners and compute the world-to-image homography matrix \( H_{wp}^p \).
     ii) Compute \( H_{i-1}^i \) for each image \( i \) and then \( H_{iw}^i \) following Equation (2.6).
     iii) Compute the camera pose \( R \) and \( t \) for each camera image following Equation (2.2).
   - **general surface**: (Section 2.2)
     i) Initialize by two-frame stereo [78].
     ii) Track the map points in subsequent images and estimate the camera poses (\( R \) and \( t \)) using the PnP algorithm [63].
     iii) Incrementally extend the map by adding keyframes if needed.
     iv) After reconstruction, select two images where the square is fully visible, mark the corners, and calibrate the scale factor.

2. Convert the camera poses to those of the 2D ultrasound images through the rigid transformation found by ultrasound calibration (Section 4.2).

---

**Figure 2-1**: Summary of the scan process and the two algorithms for simultaneous camera tracking and skin mapping.
2.1 Planar Skin Surface Mapping: Homographies

Unlike computer vision systems for general scenes, the proposed system is designed only for local skin surfaces, which are static and relatively constrained scenes for which the mapping algorithm could be optimized. In this section, the camera tracking and skin mapping algorithm based on planar homographies is described. The skin surface under the whole ultrasound scan is approximated as a planar structure, and a series of planar homographies are computed from the image sequence of skin features. The camera pose at each ultrasound image acquisition is then determined from the corresponding homography matrix.

2.1.1 World-to-Image Planar Homographies

For a perspective camera [41], the relationship between the homogeneous world coordinates \([XYZl]^T\) and the corresponding image coordinates \([x \ y \ 1]^T\) can be expressed as follows up to a scaling factor:

\[
[x \ y \ 1]^T = K [R|t] [XYZl]^T = K [r_1 \ r_2 \ r_3 |t] [XYZl]^T,
\]

where the superscript \(T\) denotes matrix transpose and \(K\) is the 3-by-3 projection matrix that incorporates the intrinsic parameters of the camera. The rotation matrix \(R\) and the translation vector \(t\) describe the geometric relationship between the world and camera coordinate systems; \(r_1, r_2\) and \(r_3\) are the column vectors of \(R\).

For points lying on a planar structure in the scene, the expression in (2.1) can be simplified without loss of generality by defining the plane as \(Z = 0\):

\[
[x \ y \ 1]^T = K [r_1 \ r_2 \ r_3 |t] [XY0 \ 1]^T = K [r_1 \ r_2 |t] [XY1]^T.
\]

Therefore, the world coordinates of points on a plane, \(X \equiv [XY1]^T\), and the corresponding image coordinates, \(x \equiv [x \ y \ 1]^T\), can be related by a 3-by-3 matrix \(H \equiv K [r_1 \ r_2 |t]\), which is usually called the planar homography matrix. Actually,
the image coordinates in different views of a planar structure are also related by a 3-by-3 homography matrix [41].

### 2.1.2 Computing Homography Matrices

For two coordinate systems that are known to be related by a 3-by-3 homography matrix (e.g., world and image coordinates of a planar structure or two views of the structure), the relationship can be expressed by the following equation in homogeneous coordinates:

\[ \mathbf{x} = H \mathbf{x}, \]  

(2.3)

where \( \mathbf{x} = [\hat{x} \ y \ 1]^T \) and \( \mathbf{x} = [x \ y \ 1]^T \) are a pair of corresponding points in the two coordinate systems. \( H \) is the homography matrix that maps point \( x \) to \( \hat{x} \).

Since \( H \) has eight degrees of freedom in the homogeneous representation, with at least four point correspondences, a linear least-square solution of the homography matrix could be found by the Direct Linear Transformation (DLT) algorithm, which could then be used as an initial estimate in an algorithm for iterative refinement [41]. Writing \( H \) as a 9-vector \( h = [h_{11} \ h_{12} \ h_{13} \ h_{21} \ h_{22} \ h_{23} \ h_{31} \ h_{32} \ h_{33}]^T \), the constraints given by \( N \) corresponding points, \( \mathbf{x}_i \) and \( \mathbf{x}_i \) for \( i = 1, 2, 3, \ldots, N \), can be written in a compact form: \( A h = 0 \), where \( A \) is a \( 2N \)-by-9 matrix:

\[
A = \begin{bmatrix}
    x_1 & y_1 & 1 & 0 & 0 & 0 & -x_1 \hat{x}_1 & -y_1 \hat{y}_1 & -\hat{x}_1 \\
    0 & 0 & 0 & x_1 & y_1 & 1 & -x_1 \hat{y}_1 & -y_1 \hat{y}_1 & -\hat{y}_1 \\
    \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
    x_N & y_N & 1 & 0 & 0 & 0 & -x_N \hat{x}_N & -y_N \hat{y}_N & -\hat{x}_N \\
    0 & 0 & 0 & x_N & y_N & 1 & -x_N \hat{y}_N & -y_N \hat{y}_N & -\hat{y}_N
\end{bmatrix}.
\]

(2.4)

Up to a scaling factor, the solution \( h \) is the unit eigenvector of the matrix \( A^T A \) with the minimum eigenvalue. Given a large number of corresponding points, the homography matrix can be robustly estimated based on random sample consensus (RANSAC) [31], which is discussed in greater details in Section 2.2.4.
2.1.3 Recovering Camera Rotation and Translation

From the above discussion, it can be seen that as long as $K$ is known and the world-to-image homography matrix is determined, the column vectors $r_1$, $r_2$, and $t$ can be found. The remaining vector $r_3$ is then simply the cross product of $r_1$ and $r_2$ due to properties of rotation matrices. In other words, for each image $i$, the camera position and orientation in the world coordinate system can be recovered, as long as $K$ and $H_i^w$, the homography matrix from the world to image $i$ coordinates, are known. Note that in practice, the recovered matrix $\hat{R}$ is rarely a proper rotation matrix, which should be orthonormal and has a determinant of 1. Therefore in this system, the nearest proper rotation matrix $R$ to $\hat{R}$ is found [43, 74], which minimizes the Frobenius norm of the matrix difference (i.e. $R - \hat{R}$).

Specifically, with the singular value decomposition (SVD) of $\hat{R}$ denoted by $U\Sigma V^T$, $R$ is equal to $UCV^T$, where $C$ is a 3-by-3 matrix:

$$C = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & \det(UV^T) \end{bmatrix}. \quad (2.5)$$

The determination of 2D-to-3D correspondences for computing $H_i^w$ usually requires defining and calibrating against a geometric pattern with known physical dimensions in image $i$, such as a square. Nevertheless, once one of these homographies is determined, $H_1^w$ for instance, $H_i^w$ for $i = 2, 3, \ldots$ can be efficiently computed in an iterative manner by finding $H_2^w, H_3^w, H_4^w, \ldots, H_{i-1}^w$ [98]:

$$H_i^w = H_{i-1}^w \ldots H_2^w H_1^w H_i^w. \quad (2.6)$$

The matrix $H_{i-1}^w$ denotes the planar homography from image ($i - 1$) to the next image, for which the determination of keypoint correspondences can be performed automatically. Note that although in this example, image 1 is used to build the connection to the world coordinates, an arbitrary selection out of the
image sequence could work as well, as long as the known geometric pattern is fully visible in that image. It should also be noted that since the final estimate $H^I_w$ is computed by cascading a series of homography estimates, in practice, errors will inevitably accumulate and drift will occur, as could be observed in Fig. 2-2(b).

As planar homographies are used for relating skin feature points, this method is particularly suitable for skin surfaces that are locally planar, such as the abdomen and breast. Under circumstances where this planarity assumption is severely violated, general skin mapping based on epipolar geometry would give better performance, as discussed in Section 2.2.

2.1.4 Experimental Results

Experiments were performed to validate the tracking and mapping algorithm based on planar homographies. The pattern chosen to provide features is shown in Fig. 2-2(a). At the top left of the pattern, there is a square with known dimensions to establish the world coordinate system in reconstruction.

In this experiment, a camera was rigidly mounted on a metal stand, as shown in Fig. 2-2(a), and was moved along the grid printed below the pattern. The camera captured a portion of the pattern at each stop in the grid, which is indicated by the blue marks. The grid dimensions were known, so the ground truth of camera translation and rotation was available for validation.

The experimental results are shown in Fig. 2-2(b). Each asterisk denotes a position where a camera image was captured. It can be seen that the estimated camera positions are very close to the ground truth. In fact, the translational error was about 1 mm on average in this case although the accumulation of drifting errors is noticeable. More evaluation of the algorithm in the context of freehand 3D US is presented in Section 5.1 and 6.1.
Figure 2-2: An experiment was performed to validate the tracking and mapping algorithm based on planar homographies. In the setup (a), the camera moved along the grid and captured a portion of the pattern at each stop. In the results (b), it can be seen that the camera position estimates (blue) are very close to the ground truth (red). The translational error was about 1 mm on average.

2.2 General Skin Surface Mapping: Visual SLAM

In this section, a camera tracking and skin mapping algorithm based on epipolar geometry is described, which is similar in concept to the monoSLAM [25] and PTAM [56] algorithms. This algorithm does not rely on any assumption on the scene, so it is applicable to skin surfaces of any size and shape. The skin features recorded in the freehand scanning process are used to incrementally map the skin surface and estimate the 6-DoF camera pose corresponding to each ultrasound image acquisition with respect to the skin map, which is a process often referred to as visual SLAM (simultaneous localization and mapping). A Bayesian probabilistic framework could be applied to fuse information from visual SLAM with ultrasound images and a prior motion model for optimal pose estimation, as described in Section 2.3. The whole reconstruction process is summarized in Fig. 2-3 and described in detail in the following sections.
extension stage
add map points
& feature matching
add frame

4
set of keyframes become the nearest keyframe
between frame pose estimation in a
Bayesian framework a
RANSAC scheme (Section 2.3)

YES
NO

4
input ultrasound video, skin feature video

output

transformation

from camera to ultrasound video

skin feature video

LOCALIZED 2D US images and 2D US coordinates calibration files

transformation

scale

end of map points

Figure 2-3: Flowchart for the visual SLAM reconstruction process.
2.2.1 Two-Frame Initialization

Visual SLAM is initialized by two camera views with a non-zero baseline distance and overlapping fields of view. Feature correspondences between the two camera frames are first established by using the SIFT method, which is relatively robust to affine distortion and change in illumination [65]. Based on these initial correspondences, the relative pose between the two views is found. This pose is widely known as the relative orientation in photogrammetry, and could be robustly estimated by applying the five-point algorithm within a random sample consensus (RANSAC) scheme [78]. In RANSAC, many random samples containing five point correspondences are taken, which yield a number of pose hypotheses. Each pose hypothesis is scored by the number of feature correspondences that are consistent with the pose by a threshold. The best hypothesis is selected as the pose estimate, and the two initial frames are defined as “keyframes” for future map extension.

An initial 3D skin map is established from the inlier feature correspondences for later tracking and probe relocalization. Each skin map point is represented by a 3-vector for its 3D position, which is computed by triangulating the corresponding pair of feature points [41], and two 128-vectors for the SIFT descriptors extracted respectively from the two camera frames.

Finally, the set of map point positions (denoted by $X$) and the set of camera poses (denoted by $R$ and $t$) are jointly optimized by bundle adjustment, where the sum of re-projection errors $E_{\text{reproj}}(X,R,t)$ is minimized [41]:

$$E_{\text{reproj}}(X,R,t) = \sum_{i=1}^{2} \sum_{k} D(x_{ik}, P(X_k, R_i, t_i)).$$  \hspace{1cm} (2.7)

The symbol $i$ indexes the images and $k$ indexes the map points. $R_i$ and $t_i$ denote the rotation and translation of the camera, respectively, at the time image $i$ is taken. (Here the pose is determined with respect to image 1, so $R_1 = I$ and $t_1 = 0$.) $X_k$ is the $k$-th map point and $P(X_k, R_i, t_i)$ represents its perspective projection onto image $i$. The function $D$ computes the distance between this projection
and the corresponding keypoint $x_{ik}$ in the frame. Since the rotations are assumed to be small, the rotation matrix $R_i$ is parameterized by the axis-and-angle representation in optimization, where the norm of the 3-vector is bounded by $\pi$. This optimization could be made significantly more efficient by utilizing sparsity in the correlation matrix for the variables [64].

Selection of the distance function $D$ is related to noise properties of keypoint localization. For instance, when only zero-mean Gaussian noise is present in the keypoint coordinates, defining $D$ as the Euclidean distance function (i.e. L-2 norm) would give the maximum-likelihood estimate:

$$D(x, \hat{x}) \equiv \sqrt{(x_1 - \hat{x}_1)^2 + (x_2 - \hat{x}_2)^2}.$$  \hspace{1cm} (2.8)

Nevertheless, this distance function is not robust to outliers in the feature correspondences. Since errors for outliers tend to be uncomonly large, they could not be accurately described by a zero-mean Gaussian model, and intuitively, the sum of distances will be inappropriately dominated by outliers.

A common distance function in the presence of outliers is the L-1 norm, which reduces the influence of outliers:

$$D(x, \hat{x}) \equiv |x_1 - \hat{x}_1| + |x_2 - \hat{x}_2|.$$  \hspace{1cm} (2.9)

A drawback of the L-1 norm is that it is not differentiable anywhere and hence could make certain optimization methods unsuitable (e.g. gradient-based optimization). An alternative to the L-1 norm could be used, which is largely similar to the L-1 norm but differentiable anywhere:

$$D(x, \hat{x}) \equiv \sqrt{(x_1 - \hat{x}_1)^2 + \epsilon^2} + \sqrt{(x_2 - \hat{x}_2)^2 + \epsilon^2}.$$  \hspace{1cm} (2.10)

The symbol $\epsilon$ represents a nearly negligible numerical quantity. In addition to the distance functions discussed above, a variety of robust statistics have been studied [50].
2.2.2 Camera Pose Tracking

After the initialization stage, subsequent camera views are localized by using the established map points as landmarks, which are associated with the SIFT feature points in each camera frame. For a current camera view, a constant-velocity motion model is employed to predict the pose based on previous pose estimates, and the map points are perspective projected to the camera frame based on the camera intrinsics. Within a pre-determined window around each projected map point, exhaustive search is performed to find the SIFT feature point in the current frame that is most probable to correspond to the map point based on descriptor similarity, which is defined by the Euclidean distance between descriptors. The window size is determined empirically based on the expected camera frame rate and probe motion.

By using these correspondences between the 3D map points and 2D feature points in the current frame, the current camera pose could be found. This task is widely known as the exterior orientation problem in photogrammetry, which is robustly solved by using an efficient Perspective-n-Point (PnP) algorithm within the RANSAC scheme [63]. Finally, this pose estimate could be optionally refined in a Bayesian probabilistic model that takes into account the re-projection errors in the camera frame, the ultrasound image regularity, and a prior motion model (Section 2.3).

2.2.3 Map Extension

In order to maintain the number of map points visible in the camera frame, new map points are added to the skin map when the currently tracked map points cover less than a certain amount of the image area (e.g. one third of the area). Feature correspondences are established by the SIFT method between the current frame and the nearest keyframe. These corresponding feature points are triangulated and the resulting 3D map points that are consistent with the pose estimates by a threshold are added to the skin map. Bundle adjustment is performed over
the current frame, the nearest three keyframes, and the map points associated with these frames. Finally, the current frame is added to the set of keyframes for future map extension. Note that the nearest keyframe is not necessarily the optimal keyframe for map extension, as too short a baseline distance between the current frame and the nearest keyframe could result in too little parallax and thus large errors in estimating map point positions. Therefore, heuristics could be designed to detect this case and to select a more suitable keyframe for map extension. It should also be noted that map extension is based on the current pose estimate, which could introduce errors in the map. Since future map extension will in turn depend on estimates based on the current map, errors will inevitably accumulate and drift will occur, as could be observed in Fig. 6-10.

2.2.4 More on RANSAC

From the above description, it could be seen that RANSAC is heavily used in the algorithm. Here I discuss how the RANSAC method could be adjusted to better fit specific applications. One of the important parameters in RANSAC is the threshold for discrimination between inliers and outliers, which is often determined empirically. Another important parameter is the number of random samples that are examined throughout the process (denoted by \( N \)), which is related to a desired probability of success (\( P \)) given the inlier ratio (\( p \)). For instance, when applying the five-point algorithm within the RANSAC scheme, a random sample consists of five feature correspondences, and hence we have:

\[
P \approx 1 - (1 - p^5)^N,
\]

which could be rewritten as:

\[
N \approx \frac{\ln(1 - p)}{\ln(1 - p^5)}.
\]
Figure 2-4: (a) Probability of success in RANSAC against the inlier ratio given varying numbers of samples. (b) Required number of samples in RANSAC against the inlier ratio given desired probabilities of success.

The two formulas are illustrated in Fig. 2-4. From Fig. 2-4(a), it can be seen that for a pre-defined number of samples, there is a threshold for the inlier ratio under which the probability of success rapidly decreases. For 100 random samples, for instance, this specific inlier ratio is roughly 0.55. From Fig. 2-4(b), it can be seen that given a desired probability of success, the required number of random samples heavily depends on the inlier ratio. Therefore, the inlier ratio of feature correspondences is important for the success application of the five-point algorithm within the RANSAC scheme. (An in-depth discussion on the inlier ratio in feature matching between natural skin features is given in Section 3.2.) Similar analysis could be performed for application of the PnP algorithm.

There are a number of variants of the RANSAC approach for specific requirements. For instance, the inlier ratio is often hard to estimate beforehand or is varying between instances. Therefore, an adaptive number of samples could be drawn in RANSAC. Given a desired probability of success, one could assume a low inlier ratio first and compute the required number of random samples. This assumption is updated after finding inliers in each iteration of the RANSAC process, which results in potential reduction of the required number of samples. The RANSAC process then stops when the existing number of samples exceeds the required number. In addition, one could perform preemptive RANSAC for more efficient
computation [79], the cost functions could be adjusted for statistically optimal performance [112], and the samples could be generated based on prior knowledge such as the descriptor similarity between feature correspondences [111, 22]. A survey and comparison of the variants could be found in [93].

2.2.5 Scale Calibration

The above visual SLAM process gives a 3D reconstruction up to a scaling factor, so the square sticker with known dimensions shown in Fig. 2-5 is affixed to the skin surface of the scan region prior to scanning to provide a reference scale. This square needs to be visible in at least two camera frames, and the four corners (marked by red circles) are manually located with sub-pixel accuracy (based on Harris corner detection [40]) in two camera views where the square is visible. The four smaller squares around the main square are designed to make corner localization more accurate and reliable. Note that in some clinical settings, this sticker might need to be sterilized.

Triangulation is performed on the identified corners to produce four 3D points, which should constitute a square in the 3D space. A 3D square with a variable pose and size is optimally fitted to the four points, and the resulting size gives a reference for conversion between dimensions in reconstruction and the real scene. Note that this procedure calibrates the scales in the camera trajectory, the skin map and the ultrasound volume at the same time since all these dimensions have been related to each other by transformations obtained from visual SLAM and ultrasound calibration (Section 4.2).

In practice, this scale calibration could be least robust when the camera optical axis is perpendicular to the plane of the square [129]. Although this configuration is not uncommon in typical use cases given the current hardware design (Section 4), the average distance between the 3D points and the fitted square has been found to be consistently less than 0.01 mm, which suggests the high accuracy of this calibration.
2.2.6 Experimental Results

Experiments were performed to validate the performance of the visual SLAM algorithm, where planar and cylindrical objects were reconstructed. The object surfaces were covered by a random pattern with a known square on the left for scale calibration, as shown in Fig. 2-6. This random pattern can be replaced by any pattern with rich features, as in the experiment described in Section 2.1.4. Part of the camera translation was measured in order to obtain the ground truth for quantitative evaluation, but the measurements were not used in tracking and mapping.

The experimental setups are shown in Fig. 2-7. In both cases, the first two images for initialization were captured 1 cm apart without rotation. Subsequently, for reconstructing the planar object, the camera moved along a straight line without rotation. For the cylindrical object, however, the object was rotated after initialization while the camera was fixed, in order to create relative motion between the camera and the object. Therefore in the cylindrical case, the expected camera motion is pure translation by 1 cm followed by rotation around the cylinder.

The reconstruction results of camera tracking and scene mapping are shown in Fig. 2-8 for both the planar and cylindrical objects. It can be seen that both the camera motion and scene maps are as expected. Specifically, in both the cases, there is only a 2%-3% error in estimating the initial 1 cm translation. More evaluation of the algorithm in the context of freehand 3D US is presented in Section 5.2, 6.2 and 6.3.
Figure 2-6: A pattern with random dots and a calibration square on the left.

(a) planar object  (b) cylindrical object

Figure 2-7: Experiments were performed to validate the visual SLAM algorithm. In the planar case (a), the camera moved along a straight line. In the cylindrical case (b), the camera moved along a straight line initially by 1 cm. Then the camera was fixed, and the cylindrical object rotated to create relative motion.

2.3 Bayesian Probabilistic Framework

For optimal pose estimation, a Bayesian framework is developed to fuse information from visual SLAM, ultrasound images, and a prior motion model. In this formulation, the 6-DoF camera pose at time instance $i$ with respect to the first pose ($i = 1$) is denoted by the 6-vector $v_i = [x_i, y_i, z_i, \alpha_i, \beta_i, \gamma_i]^T$, which includes 3-DoF translation ($x, y, z$) and 3-DoF rotation ($\alpha, \beta, \gamma$, which denote yaw, pitch, and roll, respectively). $\dot{v}_i$ denotes the temporal derivative of $v_i$. Further, $I_i$ denotes the ultrasound image at $i$, $U_i$ the set of feature points tracked in camera frame $i$, and $X$ the set of 3D map points. Here the Markov conditional independence is assumed, which means that, conditioned on the state at $(i - 1)$, the state at $i$ is independent of all the other prior states. The posterior probability of $v_i$ given the
Figure 2-8: The experimental results of the visual SLAM algorithm on (a) planar and (b) cylindrical objects, showing the scene map on the left and camera trajectory on the right. It can be seen that both results are consistent with the experimental setups.

state at time \((i - 1)\) can therefore be written as:

\[
P(v_i|v_{i-1}, \dot{v}_{i-1}, I_i, I_{i-1}, X, U_i)
\]

This representation can be simplified by conditional independence. For instance, in the first term, conditioned on \(v_{i-1}\) and \(\dot{v}_{i-1}\), \(v_i\) depends only on instantaneous acceleration and is thus independent of both \(I_{i-1}\) and \(X\). Similarly, since the current ultrasound image depends only on the previous image and their relative pose, conditioned on \(I_{i-1}\), \(v_{i-1}\) and \(v_i\), \(I_i\) is independent of \(U_i\), \(X\) and \(v_{i-1}\). Conditioned on \(v_i\) and \(X\), \(U_i\) is independent of \(I_{i-1}\), \(I_i\), \(v_{i-1}\) and \(\dot{v}_{i-1}\). Hence, the
optimal estimate \( v_i^* \) satisfies:

\[
v_i^* = \arg\max_{v_i} P(v_i|v_{i-1}, v_{i-1}) P(U_i|\mathcal{X}, v_i) P(I_i|v_{i-1}, v_i, I_{i-1}). \tag{2.14}
\]

As indicated above, three sources of information are included in this optimization: prior motion assumption, re-projection errors in camera tracking, and ultrasound image regularity, respectively. In the following sections, modeling of the three components are described in detail.

**Prior Motion Model**

A constant-velocity motion model is assumed and hence change in velocity \( \dot{v}_i \) is modeled by mutually independent zero-mean Gaussian random variables. Approximating \( \dot{v}_i \) by \( (v_i - v_{i-1}) \) and denoting the \( k \)-th element of \( \dot{v}_i \) by \( \dot{v}_{i,k} \), the estimate \( v_i \) that maximizes \( P(v_i|v_{i-1}, v_{i-1}) \) also minimizes the energy \( E_{\text{prior}}(v_i) \):

\[
E_{\text{prior}}(v_i) = \frac{1}{\sigma_t^2} \sum_{k=1}^{3} (\dot{v}_{i,k} - \dot{v}_{i-1,k})^2 + \frac{1}{\sigma_R^2} \sum_{k=4}^{6} (\dot{v}_{i,k} - \dot{v}_{i-1,k})^2. \tag{2.15}
\]

\( \sigma_t^2 \) denotes the variance of the translational component of \( \dot{v}_i \), assuming the variances are identical in the three directions. Similarly, \( \sigma_R^2 \) denotes the variance of the rotational component.

**Re-projection Errors in Visual SLAM**

Given the pose \( v_i \) and tracked map points \( \mathcal{X}_i \subseteq \mathcal{X} \), image coordinates of tracked feature points \( u_{i,k} \in U_i \) are modeled by mutually independent Gaussian random variables, the means being projections of the corresponding map points \( X_{i,k} \in \mathcal{X}_i \).

Denoting the projection of \( X_{i,k} \) to a camera frame with pose \( v_i \) by \( \text{Proj}(X_{i,k}, v_i; \theta) \), the estimate \( v_i \) that maximizes \( P(U_i|\mathcal{X}, v_i) \) should minimize the energy \( E_{\text{reproj}}(v_i) \):

\[
E_{\text{reproj}}(v_i) = \frac{1}{\sigma_u^2} \sum_{k=1}^{N_i} ||u_{i,k} - \text{Proj}(X_{i,k}, v_i; \theta)||^2, \tag{2.16}
\]
where $\theta$ denotes the set of camera intrinsic parameters and $N_i$ is the number of tracked points. Note that this form is similar to Equation (2.7) from bundle adjustment but slightly different in that the map points $X_{i,k}$ are fixed here.

Ultrasound Image Regularity

When two closely spaced ultrasound images are accurately localized, the neighboring pixels between the images should have similar intensities due to smoothness of tissue structures and limits in the spatial resolution in ultrasound imaging. Denoting the $k$-th pixel intensity in image $I_i$ by $I_{i,k}$ and its image coordinates by $p_{i,k}$, the function $F(p_{i,k}, v_i, v_{i-1}, I_{i-1})$ is defined to give the intensity of its nearest point in image $I_{i-1}$, which involves projection of $p_{i,k}$ onto the plane of $I_{i-1}$ in world coordinates and interpolation on intensities. The sum of absolute differences between all the neighboring pixel intensities is modeled by an exponential random variable with standard deviation $\sigma_i$. Hence, assuming there are $M$ pixels in the region of interest of each ultrasound image, the pose $v_i$ that maximizes $P(I_i|v_{i-1}, v_i, I_{i-1})$ also minimizes the energy $E_{us}(v_i)$:

$$E_{us}(v_i) = \frac{1}{\sigma^2} \sum_{k=1}^{M} |I_{i,k} - F(p_{i,k}, v_{i-1}, v_i, I_{i-1})|.$$  

(2.17)

In the system, only the central portion of an ultrasound image is used for this computation, as illustrated in Fig. 2-9, for two reasons. First, since compression variation is inevitable between the two consecutive images, we will want to use the portion of an ultrasound image that is less sensitive to probe compression. The central portion avoids the top and two sides of an image, where tissue is compressed the most axially and laterally, respectively. The bottom is also excluded as the signal is typically weaker.

Second, depending on the relative pose between the two images, projection of some pixels around the borders of $I_i$ might fall outside $I_{i-1}$ and thus no corresponding pixel intensities can be computed. Projecting only the central portion of an image largely reduces this border effect. For projections still outside the
borders of image $I_{i-1}$, the absolute intensity difference is defined as 255, the maximum possible value for the 8-bit gray-scale representation, to penalize extreme estimates of $v_i$.

**Energy Minimization**

From the above discussions, we can see that the optimal estimate $v_i^*$ minimizes the total energy function $E_{total}(v_i)$, where $E_{total} = E_{prior} + E_{reproj} + E_{us}$. In the implementation, the variances are determined empirically, and $v_i^*$ is found by using the Levenberg-Marquardt algorithm.

### 2.4 Relocalization

The skin map established during the freehand scan and the accompanying skin feature descriptors provide reference landmarks for the skin surface under the scan and thus allow relocalization of the probe with respect to the skin surface, which is essential for recovering loss of feature tracking during scanning [125] and is useful for spatial registration of separate 3D freehand scans in nearby regions.
Probe relocalization could be performed by matching skin feature points in a new camera frame with the map points already established during the scan and applying the PnP algorithm within the RANSAC scheme, which is a procedure similar to the tracking stage of visual SLAM described in Section 2.2. There is, however, an important difference: in the tracking stage, the current camera pose could be reliably predicted from the prior pose and therefore it can be assured that a feature point in the current frame and the projection of its corresponding map point are in close proximity to each other, but this prior pose information is unavailable for relocalization. As a result, in order to relocalize the probe, feature correspondences need to be established by matching skin feature points in the current frame against the whole set of map points, which highly relies on the distinctiveness of skin features.

The feature matching approach used in the SIFT method is applied here to establish feature correspondences and evaluate the correctness [65]. Since each skin map point is established from two camera frames, either in the initialization or extension stage of visual SLAM, each map point has two descriptors (one from each frame). Ideally for a correct 2D-to-3D feature correspondence, the feature descriptor in the current frame and the two map point descriptors should be similar to each other and dissimilar to the rest of map point descriptors. Therefore, for each feature point in the current frame, the three most similar map point descriptors are found based on the Euclidean distances between descriptors. Denoting the distances by $d_1$, $d_2$ and $d_3$, respectively, where $d_1 < d_2 < d_3$, the correspondence between this feature point and the most similar map point is accepted only when these distances meet the following criteria, where $c_1$ and $c_2$ are constant thresholds not greater than 1:

$$d_1 > c_1 \cdot d_2$$  \hspace{1cm} (2.18)

$$d_2 < c_2 \cdot d_3$$  \hspace{1cm} (2.19)

In the implementation, both $c_1$ and $c_2$ were set to 0.7. Search for the most similar descriptors is performed efficiently by the approximate nearest neighbor method,
which returns nearest neighbors with a high probability and good neighbors for the remaining cases [73].

2.5 Summary

This chapter describes two algorithms for simultaneous reconstruction of the 6-DoF camera motion and skin surface. The first algorithm is based on the assumption of scene planarity, which considers constraints in the scene geometry and works well when the local skin surface could be well-approximated as a plane. The second algorithm could be performed on general skin surfaces, but the reconstructed surface is not regularized by prior knowledge of the scene geometry. These algorithms are evaluated in vitro and the performance for freehand 3D US is examined in Chapter 5 and 6. A method for probe relocalization with respect to a previously reconstructed skin map is also introduced, which is evaluated in vivo in Section 6.3.3.

A Bayesian framework is developed to integrate information from camera images, ultrasound images, and a prior motion model to obtain optimal motion estimates. In fact, in addition to pixel intensities, further information could be extracted from ultrasound images. Specifically, one could obtain accurate in-plane motion estimates between two ultrasound images, including one rotational and two translational components, by standard correlation-based template matching, for instance. The out-of-plane motion could be estimated from speckle dissimilarities, or decorrelation, between ultrasound images. Integration of speckle decorrelation as well as other information sources into the Bayesian framework is part of the future work and is discussed in Section 8.2.

Finally, the current prototype system relies on off-line processing, which was implemented mainly in MATLAB. Nevertheless in clinical settings, it will be useful to perform real-time probe localization, which could potentially be achieved by using a more low-level programming language like C or C++, perhaps with the aid of graphics processing units (GPU).
Chapter 3

Skin Features for Tracking

Identification and extraction of skin features that are robust for tracking are essential to the proposed system. It is possible to use either artificial or natural skin features for probe localization. Artificial skin features are created in the scan region by affixing a temporary tattoo with high-contrast patterns to the skin surface, which could be made easily removable by adding an extra layer of transparent dressing between the tattoo and skin surface. Nevertheless, the use of natural skin features is often preferable since it avoids sterilization issues in clinical settings and allows probe relocalization with respect to a previously acquired skin map (Section 2.4). In this chapter, the use of both artificial and natural skin features for ultrasound probe localization is discussed and evaluated. Part of this chapter has been published in [105, 106, 107].

3.1 Artificial Skin Features

Artificial skin features are created by affixing a temporary tattoo sticker with rich features to the skin surface. A high-contrast binary random pattern is used to provide features and, as a scale reference in reconstruction, a square with known dimensions is added to the pattern (Section 2.2.5). Fig. 3-1(a) shows an example pattern where the square is 3 mm × 3 mm. Note that these random features are used only for establishing robust correspondences between camera frames,
Figure 3-1: (a) A random binary pattern as artificial skin features. A square with known dimensions is added to this pattern for scale calibration. (b) A human subject’s thigh covered by the transparent dressing and then the temporary tattoo.

and not as pre-defined fiducials for locating the camera. Therefore, design of the pattern is flexible and could be modified to optimize performance of feature detection. For instance, adding specially designed edges into the features could potentially improve the system’s robustness to motion blur, as explored in [57].

An extra thin layer of medical-grade transparent dressing (e.g. 3M Nexcare Tegaderm) could be added between the temporary tattoo and skin surface, since removing the temporary tattoo directly from the skin surface requires the use of alcohol and is thus inconvenient. As an example, Fig. 3-1(b) shows a temporary tattoo on a transparent dressing attached to the thigh of a human subject. Nevertheless, from experiments on human bodies, it was found that with the additional layers of a temporary tattoo and transparent dressing, ultrasound image brightness could be reduced. Additionally, air bubbles introduced by these layers could result in severe artifacts in ultrasound images. Therefore, even though artificial skin features guarantee robust feature tracking for pose estimation, the degradation in image quality should be considered with respect to specific clinical needs.

3.2 Enhanced Natural Skin Features

In the proposed system, it was found that, by using the camera at a short distance to the skin surface, various skin features could be observed, including uneven
distribution of melanin and hemoglobin pigments, pores, and skin surface texture [119]. These features, however, tend to be less observable in camera images when the skin surface is covered by ultrasound transmission gel, which also increases scene specularity. Water and mineral oil could be alternatives to ultrasound gel as coupling media and, partially due to the much lower viscosity, they provide better optical properties for feature extraction. Nevertheless, water tends to dry quickly and mineral oil might damage the lens materials of some transducers. Therefore, ultrasound gel is still the most suitable coupling medium and it is essential to reliably extract natural skin features from camera images even when the skin surface is covered by gel.

In order to increase the number of skin feature points that could be robustly tracked even when gel is applied, before feature detection, the recorded camera images are contrast-enhanced by contrast limited adaptive histogram equalization (CLAHE) [86]. Histogram equalization is a technique that linearizes the cumulative distribution function of pixel intensities to make the use of gray-scale levels more efficient. The CLAHE method applies histogram equalization locally to non-overlapping regions of the image and limits the contrast in each region to suppress potential enhancement of noise. Fig. 3-2 shows natural skin features at four differ-
ent body parts of a human subject, where there was evenly applied transmission gel on the skin surfaces, and it can be seen that after CLAHE, the skin features are significantly more pronounced. Nevertheless, noise in the images might have been enhanced at the same time, which could result in less stable feature detection and descriptor extraction, and thus could deteriorate performance of feature matching. Note that here color camera images (represented by RGB channels) are converted to gray-scale images by extracting the brightness channel (Y) in the YIQ color space, and later processing is performed only on the gray-scale images. Performance in feature matching with varying gray-scale conversion approaches is discussed in Section 3.2.2.

3.2.1 Lighting and Camera Exposure

Experiments were conducted on human subjects to examine the performance of matching enhanced skin features from different camera views and the influence of image signal-to-noise ratio (SNR), which is mainly determined by the lighting condition, camera exposure level, and skin tone. In the experiments, each human subject was asked to rest his or her forearm still on a table throughout the whole process and ultrasound transmission gel was evenly applied to the skin surface of the forearm. The lightweight camera described in Chapter 4 was mounted to a computer-controlled linear stage and moved for 5 cm at a constant speed to record skin features of the forearm from about 27 mm above. During scanning, the scan region was illuminated by a fluorescent lighting source obliquely incident to the skin surface so as to avoid specular reflection from gel in camera images.

For each human subject, this scanning process was repeated four times at the same starting and ending positions under the same lighting condition with four different camera exposure levels denoted by 1, 2, 3 and 4, respectively, where level 1 gave the shortest exposure time and level 4 the longest. Between two consecutive exposure levels, the exposure time differs by a factor of two. From the video of skin features in each linear scan, eleven frames were extracted with
The performance of matching contrast-enhanced skin features between two camera views was examined on the forearms of four human subjects with varying camera exposure levels and thus varying image SNRs. (a) Skin tones of the human subjects at the forearms under the same lighting condition with exposure level 3. (b) The same forearm region of subject A under the same lighting condition with four different exposure levels. Level 1 gave the shortest exposure time and level 4 the longest. (The images shown here are not contrast-enhanced.)

an equal spacing of 5 mm, which gave ten pairs of equally spaced frames. The SIFT feature points were extracted from the frames and matched between each frame pair.

For discrimination between correct and incorrect feature correspondences, the five-point algorithm was performed between each frame pair within the RANSAC scheme with an inlier threshold of 2 pixels, which is a procedure similar to the initialization stage of visual SLAM (Section 2.2). In this procedure, a larger number of inlier correspondences indicates higher recognizability of skin features between camera views and hence more robust pose estimation. The ratio between the number of inlier correspondences and that of all SIFT feature correspondences was also examined, which describes the overall quality of feature matching. This inlier ratio also determines the required number of random samples in RANSAC to achieve a given probability of success (Section 2.2.4) and is therefore critical to computational efficiency. These two performance metrics for feature matching were averaged over the ten frame pairs and summarized in Fig. 3-4.

This experiment was performed on four human subjects, who are of different ethnicities (African American, Asian Indian, Caucasian, and Chinese) and have
Figure 3-4: Performance of matching contrast-enhanced skin features between two camera views on four human subjects with varying camera exposure levels, as shown in Fig. 3-3. (a) Number of all SIFT feature correspondences. The error bars indicate standard deviations over ten frame pairs. (b) Number of inlier feature correspondences. (c) Inlier ratios.

different skin tones. Fig. 3-3(a) shows skin tones of the human subjects at the forearms under the same lighting condition with exposure level 3, and Fig. 3-3(b) shows the same forearm region of one of the subjects under the same lighting condition with varying exposure levels.

As can be seen in Fig. 3-4(b) and (c), there were generally hundreds of skin feature points that could be identified from two different camera views, and for most cases, the average inlier ratio was above 0.4. This ratio meant that less than 670 random samples were required in RANSAC to achieve a 99.9% probability, for instance, of sampling five correct feature correspondences at least once for the five-point algorithm. It could also be seen that, with increased exposure time and hence an increased SNR in camera images, the number of inlier correspondences and inlier ratio generally improve. Although too much exposure could result in intensity saturation in camera images and the number of feature correspondences could decrease as a result, the inlier ratios were unaffected by this saturation since they reflect the quality of feature correspondences and not the quantity.

Overall, these results show the applicability of skin-based camera tracking to different skin tones, although the performance could vary with skin conditions and the image SNR. A higher SNR could be achieved by either a longer camera exposure, stronger lighting, or both. In practice, it is generally preferred to use
an external lighting source to provide stronger lighting, since a longer camera exposure is more likely to bring motion blur in camera images and hence to limit the speed of probe motion. An external lighting source is more suitable also because clinically, ultrasound imaging is usually performed in a darkened room for better perception and contrast of the acquired images. Therefore, it is challenging to rely on only the ambient light in clinical settings. The possibility of using an add-on lighting source is discussed in Section 4.3.

3.2.2 Color Space

The experimental results described above are based only on the brightness channel (Y) from the YIQ color space, but other approaches for conversion from color images (represented by RGB channels) to gray-scale images have also been evaluated. One could simply use one of the RGB channels as a gray-scale image. One could also use one of the YIQ color channels, which are designed to optimize human perception in the NTSC (National Television System Committee) color television system and are computed by taking linear combinations of the RGB channels. The Y channel represents the brightness information, and the I and Q channels represent the chrominance information. Finally, the HSV color space was also tested, which stands for hue, saturation, and value, respectively. The H and S channels represent the chrominance information, and the V channel represents the brightness. These channels are computed by taking non-linear combinations of the RGB channels.

The same experiment procedure as in Section 3.2.1 was followed to examine performance on nine different color channels as described above: RGB, YIQ, and HSV. The images under camera exposure level 3 for the four human subjects were used, and the color channels from examples of these images are shown in Fig. 3-5. The numbers of feature correspondences and the inlier ratios were computed as performance metrics for feature matching.

The experimental results are shown in Fig. 3-6. It could be seen that the change
Figure 3-5: Different color channels (i.e., RGB, YIQ, and HSV) from the skin images of the four human subjects shown in Figure 3-3(a).
Figure 3-6: The performance of matching contrast-enhanced skin features between two camera views was examined on the forearms of four human subjects by using different channels in three color representations: RGB, YIQ, and HSV. The subjects were the same as shown in Fig. 3-3 and the camera exposure was fixed at level 3. (a) Number of all SIFT feature correspondences. The error bars indicate standard deviations over ten frame pairs. (b) Number of inlier feature correspondences. (c) Inlier ratios.

In performance among color channels is somewhat different between subjects, which is expected since the subjects have different skin tones. Nevertheless, a number of observations could be drawn from Fig. 3-6. First of all, among the RGB channels, the G channel gives the best performance, which should be due to the fact that many skin features are resulted from uneven distribution of melanin pigments. One could also see that generally, the channels representing only chrominance, such as I, Q, H, and S, do not perform as well as the other channels. This observation suggests that the skin features due to change in brightness tend to be more reliably matched. Finally, the Y channel gives significantly better performance than all the other color channels tested. Therefore, the Y channel is used in the proposed system to obtain gray-scale images for further processing. Note that since image brightness was found to give the most relevant information for feature extraction, imagers without color filters could potentially give even better results. Also, this experiment was performed by using the SIFT feature extraction method on single color channels, and different feature descriptors, especially those that take into account color information by design, could give different results.
3.3 Summary

This chapter discusses the use of both artificial and natural skin features for scene reconstruction and camera motion estimation. Artificial skin features are high-contrast, which allow robust tracking but are less suitable for clinical uses. In contrast, the use of natural skin features is convenient, does not require additional sterilization, and allows probe relocalization over time.

The performance of matching natural skin features extracted from contrast-enhanced camera images is evaluated on varying camera exposures, skin tones, and color conversion schemes. It was found that the performance is optimal when skin features are extracted from gray-scale images representing the brightness channel (i.e. Y channel in the YIQ color model), and the performance improves with a longer camera exposure or stronger lighting. In fact, it is also possible to use signal beyond the visible light spectrum, such as the infrared or ultraviolet spectrum, to perform scanning in a visually darkened room. This possibility and general lighting control during scanning are further discussed in Section 4.3.

In addition to the issues already discussed, a number of challenges exist in tracking natural skin features. For instance, the presence of hair could give rise to more features for tracking, but these features could be unstable between camera frames if the hair is changed by probe contact. Therefore, in the presence of hair, it is generally recommendable to complete the scan with only one sweep as the hair pattern could be different in the second sweep. Another challenging scenario for probe localization based on skin features is scanning elderly patients, who tend to have wrinkled and loose skin. In this case, the skin features are somewhat spatially decoupled with underlying tissue structures and hence inconsistency in reconstruction could be introduced. This issue remains even if artificial skin features are applied.
Chapter 4

System Hardware

This chapter presents hardware design for the probe localization system, including the camera, probe housing, and lighting source. Design and calibration of the hardware are first described. The use of a lighting source is then discussed. Finally, the issue of skin surface deformation due to probe contact is investigated. This issue is addressed by hardware design, which is validated through experiments on the human body. Part of this chapter has been published in [104, 105, 106, 107].

4.1 Camera and Probe Housing\(^1\)

In the system, the Terason t3000 ultrasound imaging system (Terason Ultrasound, Burlington, MA, USA) is used with a 5-MHz linear array transducer, and an off-the-shelf USB web camera (Macally IceCam2) is mounted to the probe for recording skin features. In addition to being small, lightweight and low-cost, this camera allows manual focus adjustment and a focus distance as short as 2 cm, which is particularly suitable for the proposed system.

A plastic probe housing was designed and 3D-printed to rigidly hold the camera as shown in Fig. 4-1, which is composed of two parts: one is attached to the

\(^{1}\)The probe housing was designed and 3D-printed with the help of Matthew W. Gilbertson in the Department of Mechanical Engineering at MIT.
Figure 4-1: The ultrasound probe with a rigidly mounted camera for recording skin features.

Figure 4-2: The checker-board pattern for camera calibration. The square size is 12×12 pixel, which amounts to around 0.5×0.5 mm when printed with 600 dots per inch (DPI). The white dot at the top left marks the origin of the coordinate system in calibration.

probe, and the other holds the camera. With the aid of 3D surface models from laser scanning, the housing on the probe tightly fits the surface shape, which prevents relative motion between the housing and the probe. The camera is mounted to the probe housing by using magnets with kinematic coupling, which allows easy and repeatable attachment. Therefore, as long as the probe housing remains fixed on the probe, ultrasound calibration (as described in Section 4.2) only has to be performed once even if the camera needs to be detached frequently.

The camera was focused at the skin surface from about 27 mm above and has a 640×480 spatial resolution, which maps to an approximately 28 mm × 21 mm field of view (FOV) on the patient's skin surface when there is little probe compression. The camera center is about 4 cm from the surface of the probe such that local skin surface deformation due to probe pressure is mostly avoided.
in the FOV (Section 4.4). During scanning, camera frames and 2D ultrasound images are acquired synchronously at around 10 frames per second and stored as video files through lossless compression (Huffyuv video codec) for off-line reconstruction. On average, the latency between ultrasound and camera image acquisition is less than 50 millisecond. Given a probe travel speed of around 0.5 cm/s in freehand scanning, for instance, the probe travel distance during this latency is less than 250 μm. Since the elevational resolution of ultrasound imaging at a center frequency of 5 MHz is normally around 1 mm [35, 94], the change in ultrasound images caused by a separation distance of 250 μm is insignificant. Therefore, the ultrasound and camera images are considered to be acquired at the same time instance.

Through the standard camera calibration procedure using the checker-board pattern shown in Fig. 4-2 [16], the camera intrinsic parameters and radial lens distortion coefficients were estimated. Around 40 camera images were captured at varying camera views in this calibration to provide sufficient constraints for the solution. The pattern was placed at different parts of the FOV to sufficiently sample the effects of lens distortion. The intrinsic parameters, including the focal lengths \( f_x \) and \( f_y \) and principal points \( (c_x, c_y) \), determine the homogeneous projection matrix \( K \) for the perspective projection camera model:

\[
K = \begin{bmatrix}
  f_x & 0 & c_x \\
  0 & f_y & c_y \\
  0 & 0 & 1
\end{bmatrix}.
\]

The radial lens distortion coefficients describe camera image distortion due to lens geometry. Note that it is possible but unnecessary to create undistorted images before further processing. Instead, lens undistortion and distortion could be applied at only detected keypoints (e.g. SIFT feature points [65]) during the reconstruction process (Chapter 2).

One limitation of this system is the defocus blur resulted from a large tilt angle, which could not be completely addressed even with carefully designed
optics. Since the camera is focused at the skin surface from only 2-3 cm above, the depth of field is relatively shallow. It was found that the acceptable tilt angle is approximately ±20°, beyond which skin features could become too blurred to be extracted.

Transformation from the ultrasound image coordinate system to the camera is estimated by using the single-wall method and remains the same throughout the scanning process [89]. The methods for this ultrasound calibration are described in detail in the following section.

4.2 Ultrasound Calibration

Ultrasound calibration is a procedure for finding the transformation matrix that translates the estimated pose of a sensor rigidly mounted to an ultrasound probe into the corresponding probe pose. As illustrated in Fig. 4-3(a), the goal of ultrasound calibration is to determine $T_{\text{scan}}^{\text{camera}}$, the transformation from the camera coordinates $(x, y, z)$ to the ultrasound image coordinates $(u, v, 0)$, which is constant throughout the scanning process since the camera mounting is rigid. Since this transformation matrix essentially describes a rigid transformation in the 3D space, there are six degrees of freedom in this matrix (three for translation and three for rotation).

A variety of ultrasound calibration methods have been studied for systems using optical or electromagnetic trackers [48]. However, none of the existing methods could be directly applied to probe tracking using cameras as in the proposed system. Here two novel ultrasound calibration procedures are proposed for systems that involve one or more cameras rigidly mounted to an ultrasound probe. One procedure relies on alignment of the ultrasound image plane with a 2D target with a known shape and position, which is termed the "2D-target method". In the other procedure, measurements of a flat bottom in a water tank from multiple ultrasound images are combined in a single optimization framework, which determines the optimal values for the six unknowns that describe the camera-to-
4.2.1 The 2D-Target Method

In the 2D-target calibration method, the probe is fixed at a known position and orientation in a world coordinate system. This task can be performed with a sub-millimeter spatial resolution by aligning the ultrasound image plane with a known target precisely located at that position, as shown in the phantom design in Fig. 4-3(b). By this particular positioning, $T_{\text{scan}}^{\text{world}}$, the transformation from the world to the ultrasound image coordinates, can be found. The camera pose at this particular positioning can be estimated by using a checker-board camera calibration pattern placed at a known position in the world coordinate system. Therefore, $T_{\text{camera}}^{\text{world}}$, the transformation from the camera to the world coordinates, can be estimated. Combining the two transformations above, the desired transformation $T_{\text{scan}}^{\text{camera}}$ can be found:

$$T_{\text{scan}}^{\text{camera}} = T_{\text{scan}}^{\text{world}} \cdot T_{\text{world}}^{\text{camera}} \quad (4.1)$$

Fig. 4-4 shows the ultrasound calibration phantom made of metal structures and agar-based material with a camera calibration pattern.
Performance of the 2D-target method heavily relies on the alignment between the ultrasound image plane and the known target. However, it is difficult to judge quality of the alignment, and hence quality of the calibration, from a single image. Furthermore, the extracted parameters from this calibration method could be unstable since only a single image is used for calibration and measurements from multiple images could not be naturally combined to determine a single solution. Finally, preparation of the calibration phantom requires a fair amount of effort in hardware design and calibration. These limitations are addressed in the following “single-wall method”.

4.2.2 The Single-Wall Method

Due to limitations in the above 2D-target method, ultrasound calibration in the proposed system was performed by the single-wall method described in this section, which is similar to the single-wall method described in [89]. The phantom structure used in the single-wall method is a water tank with a flat bottom and a stand with a known depth $d$, as illustrated in Fig. 4-5(a). A checker-board camera calibration pattern is attached on the top of the stand. The world coordinate system $(X, Y, Z)$ in this calibration is defined by this calibration pattern, and through
extrinsic camera calibration, camera poses in this coordinate system could be determined as long as the pattern is visible in the camera FOV.

During calibration, a number of ultrasound images showing the flat bottom are acquired at varying probe positions and orientations. Along with acquisition of an ultrasound image, a camera image is captured at the same time, in which the calibration pattern should be visible. These probe poses need to be diverse to provide sufficient constraints for determining a good solution, as explained in [89]. An example ultrasound image is shown in Fig. 4-5(b). Since the flat bottom gives strong ultrasound wave reflection and hence is shown in the image as a particularly bright line, this line could be automatically extracted [89]. Note that before calibration, the flat bottom needs to be roughened, e.g. by using fine sandpaper, to improve ultrasound wave reflection at oblique angles [117].

The optimal solution of the calibration parameters is obtained by minimization of a cost function describing constraints in the extracted lines in ultrasound images. As indicated in Fig. 4-5(a), the ultrasound image coordinate system is denoted by \((u, v, 0)\) (the image plane is assumed to be \(Z = 0\) without loss of
generality.) Denoting the ultrasound-to-camera transformation by $T_{\text{camera}}$ and the camera-to-world transformation by $T_{\text{world}}$, we have:

$$[X \ Y \ d \ 1]^T = T_{\text{world}} T_{\text{camera}} [u \ v \ 0 \ 1]^T. \quad (4.2)$$

At each ultrasound image acquisition, $T_{\text{world}}$ could be different and is determined by extrinsic camera calibration using the checker-board pattern. $T_{\text{camera}}$ is the rigid transformation matrix that one wishes to find through ultrasound calibration, which remains the same throughout the whole calibration process since attachment between the camera and the probe is rigid. Note that since intrinsic camera calibration could be jointly performed with extrinsic calibration, this single-wall method allows complete camera calibration and ultrasound calibration in a single framework.

For a given point $[u \ v \ 0 \ 1]^T$ in the ultrasound image plane, although it is hard to measure the resulting $X$ and $Y$ in world coordinates, the $Z$ component should always be the known depth $d$. Therefore in an ultrasound image, a point provides a linear constraint. In the proposed system, two points near the two ends of an extracted line were used, and around 40 pairs of ultrasound and camera images were acquired for calibration. Note that here the values of $u$ and $v$ are in physical dimensions, such as centimeters or millimeters. Computation of these values highly relies on the estimate of pixel spacings, which is a function of water temperature and the resulting speed of sound [11].

If the matrix $T_{\text{camera}}$ is parameterized by 12 unknowns (9 for the rotation matrix and 3 for the translational vector), a least-square solution could be found by stacking all the linear constraints in a linear system. However, this is over-parameterization since the rotation could be described by only 3 unknowns. As a result, the least-square solution needs to be further regularized to guarantee that a valid rotation matrix is given. This regularization could be avoided if the rotation is consistently parameterized by 3 unknowns (e.g. by using the axis-angle representation) in the optimization. In this case, the constraints are non-linear in the
unknowns, so it is not straightforward to find an analytical solution. Nevertheless, a least-square solution could still be found by applying iterative optimization routines, such as the Gauss-Newton algorithm and the Levenberg-Marquardt algorithm.

4.3 Lighting

It is worth noting that the idea of adding a lightweight lighting source to the probe housing, as shown in Fig. 4-6, has been tested. The lighting source (Mighty Bright, Santa Barbara, CA, USA) is composed of two light-emitting diodes (LED) and precision optical grade lens for light diffusion. Although this design helps to make the system more compact by eliminating the need for an external lighting source and hence more usable in varying lighting conditions, it was found to be challenging to illuminate the skin surface evenly in the camera FOV. As a result, this add-on lighting source almost always causes local over-exposure or under-exposure.

This phenomenon is shown in Fig. 4-7, where the skin surface was recorded under no light (i.e. only ambient light), weak light (one LED turned on), and strong light (two LEDs turned on). It can be seen that when illumination is provided by the lighting source in addition to ambient light, there is over-exposure near the bottom of the images due to the illumination. This over-exposure causes shrinkage of the effective FOV and could not be eliminated by change of the
Figure 4-7: Skin features recorded by the camera shown in Fig. 4-6 before (top) and after (bottom) contrast enhancement, as described in Section 3.2: (a) no light, (b) weak light, and (c) strong light. (The camera exposure time in (a) is four times that in (b) and (c).)

Figure 4-8: Illustration of specular reflection from the skin surface covered by ultrasound gel. Note that when the incident angle of light is small with respect to the skin surface normal, the reflected light could enter the camera aperture, which results in specular reflection.

Therefore, in the final design as shown in Fig. 4-1, the add-on lighting source was removed, and scanning is performed by using ambient light or an external lighting source.

In contrast to the add-on lighting source, an external lighting source could provide an adjustable incident angle with respect to the skin surface, which is helpful in reducing specular reflection due to ultrasound transmission gel. As
illustrated in Fig. 4-8, the incident angle of light from an add-on lighting source is small and constrained by the probe housing. Therefore, the reflected light from the skin surface could result in specular reflection in the camera FOV. Nevertheless with an external lighting source, the incident angle could be large, and hence the specular reflection could be reduced. Additionally, since the distance between an external lighting source and the skin surface could be significantly longer than an add-on source, it is also easier to create uniform illumination in the camera FOV by using external lighting.

Finally, it is possible to record skin features beyond the visible light spectrum. For instance, by removing the infrared (IR) filter in the camera, signal in the IR light spectrum (i.e. wavelength between 700 nm and 1 mm) could be captured by the camera sensor. Since IR light generally penetrates deeper into skin tissue structures than visible light, the signal from IR light was expected to provide additional skin features such as vascular structures [67]. Nevertheless, it was found that the increase in skin features was not as significant as expected. An
example is shown in Fig. 4-9, where the image with IR light looks “redder” since the IR signal mainly passes through the R (red) portion of the Bayer filter in the camera. Although vessels are slightly more obvious in the camera images when signal from the IR spectrum is reserved, the difference is not significant.

It should be noted that this experiment was performed by using a low-cost web camera mainly for recording visible light. Vessel structures would become more obvious in images if a dedicated IR camera and an IR lighting source are used. Further, it is possible to apply ultraviolet photography to make superficial skin features much more emphasized in images [33], although there could be concerns about the influence of ultraviolet light on health of skin tissues. The use of signal beyond the visible light spectrum also improves usability of the system in clinical settings since ultrasound imaging is often performed in a purposefully darkened room for better visualization and perception.

4.4 Robustness to Skin Surface Deformation

In the camera tracking and skin mapping algorithms, a skin map is incrementally built based on the assumption that the skin surface under a scan undergoes only rigid transformation throughout the process. It appears that this assumption could be violated by probe compression, which locally deforms the skin surface of soft body parts. Nevertheless, it was found that with a carefully designed hardware configuration, there could be little skin surface distortion in the camera FOV from probe compression and hence skin mapping could essentially be performed on the undeformed portion of the skin surface at each time instance.

This avoidance of skin surface deformation in the camera FOV is illustrated in Fig. 4-10. The distance between the FOV and probe contact with no probe compression is denoted by $s_2$, which is determined by the vertical camera viewing angle, distance between the camera and skin surface $h$, and camera-to-probe distance $s_1$. (These measurements could be obtained based on the ultrasound-to-camera transformation matrix from ultrasound calibration.) To some extent, this
Figure 4-10: Illustration of skin surface deformation due to probe compression. Distance measurements before (left) and after (right) probe compression are indicated. Note that the distance between the camera FOV and probe contact increases with probe compression (hence $s_2' > s_2$).

distance $s_2$ prevents skin surface deformation from being visible in the FOV. Further, with an increased compression level, as illustrated on the right of Fig. 4-10, $h$ decreases to $h'$ and hence $s_2$ increases to $s_2'$. In other words, although harder probe compression causes more severe skin surface deformation, it brings the FOV farther away from probe contact at the same time. Therefore, even when there is significant probe compression, the locally deformed portion of the surface could be mostly avoided in the FOV as long as the hardware configuration is properly designed based on expected skin surface deformation under typical scans.

The current design ($h = 27.5$ mm, $s_1 = 42.6$ mm, $s_2 = 32.1$ mm, and camera viewing angle around $42^\circ$) was found to be working well in terms of avoiding surface deformation in the camera FOV. This design was validated by experiments performed on the abdomen of a human subject, which is one of the softest body parts. In order to ensure that there were rich features throughout the camera FOV, artificial skin features were created in the region of interest (Section 3.1). In the experiment, a camera image was first taken with minimal probe compression on the abdomen. The probe was then slightly translated (by around 4.5 mm)
to a new position, and camera images of artificial skin features were then taken at varying compression levels. Between the pre-translated image and each one of the post-translated images, the SIFT feature correspondences were found and the five-point algorithm was performed within the RANSAC scheme with a 2-pixel threshold, which is a procedure similar to the initialization stage in visual SLAM (Section 2.2). Note that since the five-point algorithm solves the relative orientation problem 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. 4-11. For each post-translated image, the compression distance against the skin surface is indicated at the top left, which was estimated based on prior knowledge of the artificial patterns in the FOV. The inlier feature points are marked in red, which are absent in the leftmost portion of post-translated images since this portion is not covered by the pre-translated FOV. As can be seen, with an increased compression level, the number of inlier features decreases due to a smaller FOV on the skin surface. Nevertheless, for all the compression levels, inlier feature points are present all over the FOV except the leftmost portion. This observation suggests
that the skin features throughout the FOVs underwent little non-rigid deformation from probe compression, including those features relatively close to probe contact (near the bottom of images), and it is therefore confirmed that local surface deformation is avoided in the camera FOV by this hardware design even with an increasing compression level.

One design possibility to make the hardware less obtrusive is to shorten the camera-to-probe distance $s_1$, which would potentially introduce surface deformation in the camera FOV and hence decrease the number of inlier features. Although the deformed skin features would be rejected as outliers in RANSAC and thus do not lead to catastrophic reconstruction failure, tracking performance could be compromised by this design change due to the decreased number of inliers available for pose estimation. This skin surface deformation might affect discrimination between rotation and translation from camera images, since part of the FOV will be unusable and hence only local visual cues are available from this essentially reduced FOV. If shrinkage of the hardware is necessary, it is possible to address the issue of deformed skin surfaces in a non-rigid framework [113].

It should be noted that, in addition to the compression force discussed above, shear force along the skin surface could cause movement of skin features with respect to the underlying tissue structure, which degrades quality of reconstruction. However, this effect is not considered in this validation since clinically, little shear force remains with the application of ultrasound gel. Another potential issue is the movement of organs with respect to skin features due to probe compression, which could cause spatial inconsistency in the reconstruction. This issue is challenging and common for all probe localization methods in freehand 3D US.

4.5 Summary

This chapter describes hardware design and calibration for the probe localization system, and a number of design considerations are discussed, including lighting and skin surface deformation due to probe contact. It was found that an external
lighting source is generally more suitable than an add-on lighting source in terms of creating uniform lighting and avoiding specularity due to ultrasound gel. Also, it is shown that although the skin surface could be deformed due to probe contact, by careful hardware design, motion estimation and surface reconstruction could still be performed only based on the undeformed portion of the skin surface.

Note that the focal length of the camera described in this chapter is manually adjusted, but most off-the-shelf web cameras feature automatic focusing for more convenient daily uses. Although it is still possible to estimate the focal length of an auto-focusing camera at a given time instance from customized software or generic EXIF tags, the estimated value is generally less accurate than what could be obtained by camera calibration on a fixed-focus camera. Further, it is also possible to perform self-calibration purely based on camera images from scanning, and not images of the checker-board pattern, to extract the intrinsic parameters [87, 34, 128], but the estimates also tend to be less accurate than conventional calibration.

Finally, the camera currently used in the system is a low-end web camera off the shelf, but it is possible to design one's own optics to build a camera that is more compact and better suits requirements of this system. For instance, one could use optics that provide a wider camera viewing angle than the angle normally featured in an off-the-shelf webcam (the viewing angle of the current camera is around 42°.) A wider viewing angle provides a larger FOV, in which there are more depth variation and more features for tracking. Nevertheless, given a fixed number of pixels and pixel spacing, a wider viewing angle also gives a poorer angular resolution and more severe lens distortion. Therefore, there is a trade-off in choosing the optimal viewing angle for camera tracking, although the results reported in [24] have suggested that a wider viewing angle improves the performance of visual SLAM. Performance of the system could also be improved by using a camera with a higher frame rate, which is helpful for reduction of motion blur and therefore allows more rapid probe motion. Other potentially helpful design choices include the use of polarizers to reduce the influence of

84
gel specularity and incorporation of programmable lighting sources, such as variable add-on lighting via feedback control for reduction of intensity saturation in camera images.
Chapter 5

In-Vitro 3D Ultrasound Experiments

This chapter presents in-vitro freehand 3D US experiments on agar-based phantoms with both planar and curved surfaces. Artificial features were created on surfaces of the phantoms for probe localization. Cylindrical inclusions with known dimensions were embedded inside the phantoms and were scanned following varying scan trajectories for quantitative evaluation of the system performance. Part of this chapter has been published in [104, 105].

5.1 Planar Surface

5.1.1 Experimental Setup

Experiments of 3D US based on planar homographies (Section 2.1) were performed on an agar-based phantom with artificial surface texture, which is shown in Fig. 5-1(a) and similar to the one created in [88]. Graphite powder was added in the phantom material to control scattering characteristics. A cylinder made from pure agar, whose volume is 1.26 mL, was embedded in the phantom for 3D US scanning. Dimensions of the phantom are shown in Fig. 5-2(a). The pattern shown in Fig. 5-1(b) was placed on the phantom surface for the establishment of a world coordinate system. An example ultrasound image of the phantom is shown in Fig. 5-2(b), where the difference in image brightness due to varying
concentration of graphite power could be seen.

Three freehand scanning types were tested on this phantom: linear, tilt, and rotational scanning. In linear scanning, the ultrasound image plane remains approximately perpendicular to the phantom surface and the probe moves along the elevational direction. In tilt and rotational scanning, only pitch and yaw of the probe is varied, respectively, as illustrated in Fig. 5-3(a).

5.1.2 Experimental Results

After the ultrasound images were acquired and their poses were estimated from the camera images, 3D volume reconstruction was performed using Stradwin [117]. The cylinder was segmented in each ultrasound image by thresholding. The cylindrical volume was then reconstructed and surface rendered. Typical scanning trajectories and reconstructed cylinders are shown in Fig. 5-3(b)-(d).

An average of twenty 2D images were acquired for each 3D volume reconstruction. Each type of scanning (linear, tilt, rotational) was performed five times. The volumes of the reconstructed cylinders were estimated in Stradwin. Fig. 5-4 shows the mean volume errors for each scanning type. The error bar represents
one standard deviation. It can be seen that linear scanning generally gave volume estimates with less than 3% error. The errors in tilt and rotational scanning were about 3.5 times larger than in linear scanning.

Linear scanning gave the smallest volume error. This observation is not surprising, since recovery of rotation from camera images is a process that is relatively sensitive to errors in feature extraction. Pitch or yaw variation can be misinterpreted as translation given a limited camera FOV, especially when there is a lack of local features or presence of camera defocus in tilt scanning. For example, this phenomenon can be seen from the z-axis misalignment between the 2D images in Fig. 5-3(c) and the cylinder shape distortion in Fig. 5-3(d), and was found to result in volume underestimation as shown in Fig. 5-4.

A major source of error in this method is drift, since once the calibration landmark is no longer in the camera FOV, probe localization relies solely on surface features and transformation back to the calibration image. Drift can be reduced by registering the square landmark whenever it is fully visible, or by building a global map of the surface features, as in the visual SLAM algorithm (Section 2.2). It is also possible to place multiple landmarks with a known spatial configuration in the scanned region in order to register the camera globally and correct drift.
Figure 5-3: (a) Illustration of the three scanning types: linear, tilt, and rotational scanning. (b)-(d) Examples of the scan trajectories and reconstructed cylinders.

Figure 5-4: Comparison of mean volume errors.

5.2 General Surface

5.2.1 Experimental Setup

Experiments of 3D US based on visual SLAM (Section 2.2) were also performed on agar-based phantoms in which cylinders with known dimensions were embedded. The cylinders were 1.1 cm in diameter and 1.7 cm in length, which gave a volume of about 1.62 cm$^3$. In order to add artificial features to the phantom surfaces, random patterns similar to the one shown in Fig. 3-1 were printed on temporary tattoo stickers. The phantom surfaces were wrapped by plastic food
Figure 5-5: (a) The phantoms with planar (top) and curved (bottom) surfaces, covered by temporary tattoo stickers. (b) Estimated surface feature points and camera trajectory (shown in blue) from scanning the phantom in (a). Part of the corresponding ultrasound image trajectory is shown at the bottom right (in gray). (c) Ultrasound scan trajectory and the reconstructed cylinder, which is highlighted in the inset.

wrap, to which the tattoo stickers were then affixed. Paste was used to avoid sliding between the phantoms and the food wrap. Finally, standard ultrasound gel was applied evenly on the surfaces as the coupling medium.

Two phantoms were made following the above procedure: one with a planar surface and one with a curved surface, as shown in Fig. 5-5(a). The embedded cylinders were scanned in freehand with roughly piecewise-linear scan paths.

### 5.2.2 Experimental Results

From the recorded camera videos, the phantom surfaces were mapped and the probe poses were estimated, as shown in Fig. 5-5(b). The cylinders in these spatially registered ultrasound images were then manually segmented in Stradwin [117]. The surface-rendered cylinders were shown in Fig. 5-5(c), along with the
ultrasound scan trajectories. Volumes of the reconstructed cylinders were estimated in Stradwin and compared with the ground truth. On average, the volume error was +2.65 % for the planar phantom and +6.30 % for the curved phantom. These results demonstrate the applicability of the visual SLAM algorithm on both planar and non-planar surfaces.

5.3 Summary

This chapter presents in-vitro experiments to evaluate performance of the two probe localization algorithms in freehand 3D US, both of which are described in detail in Chapter 2. The applicability of the visual SLAM algorithm on both planar and curved surfaces is demonstrated. Additionally, volume reconstruction errors are quantified, which could be varying with scan types and surface geometry. Further experiments on human subjects are described in Chapter 6 to quantify the 6-DoF motion errors and to examine the quality of tissue volume reconstruction.
Chapter 6

In-Vivo 3D Ultrasound Experiments

In this chapter, performance of the probe localization system is evaluated on multiple human body parts and volume reconstructions are demonstrated. The algorithms for both planar and general surfaces (Chapter 2) are validated on both artificial and natural skin features. Motion estimation errors in probe localization and relocalization are quantified. Finally, the volume reconstructions and synthesized reslices are validated by comparison with real ultrasound images. Part of this chapter has been published in [105, 107].

6.1 Nearly Planar Surface with Natural Skin Features

In this experiment, the algorithm based on planar homographies (Section 2.1) was applied to a nearly planar skin surface. Fig. 6-1 shows the in-vivo experimental setup. The world coordinate system was established by affixing a 3 mm x 3 mm square to the skin surface and manually marking the four corners in the camera image. At each acquisition of an ultrasound image, a camera image of natural skin features was recorded at the same time. Here water was used as the ultrasound coupling medium.

Fig. 6-2 shows the results of probe tracking and 3D volume reconstruction, where the direction of probe movement was approximately perpendicular to the ultrasound image plane. Fig. 6-2(a) shows the skin features aligned in the world
Figure 6-1: *In-vivo* experimental setup on a nearly planar skin surface.

Figure 6-2: *In-vivo* experimental results based on planar homographies. (a) Aligned skin features and the tenth camera pose as an example. (b) Acquired ultrasound images registered in the world coordinate system.

coordinate system, which was established by the black square at the bottom left. The camera pose for each patch of skin features was estimated, but only the camera for the tenth patch is shown here as an example. The acquired ultrasound
images were aligned in the world coordinate system according to the pose estimates, as shown in Fig. 6-2(b).

The acquired data were further processed using Stradwin [117]. The brachial artery was manually segmented in each ultrasound image and surface-rendered. In Fig. 6-3(a), an example contour of the segmented brachial artery is shown along with the ultrasound image. The rendered surface is shown in Fig. 6-3(b), where the ultrasound image in Fig. 6-3(a) and the pose of each image are also shown.

### 6.2 General Skin Surface with Artificial Skin Features

In this experiment, the visual SLAM algorithm for general skin mapping (Section 2.2) was applied to the skin surface with artificial skin features. The random pattern shown in Fig. 2-6 was printed on a temporary tattoo sticker and affixed to the human subject’s forearm, which was to be scanned, as shown in Fig. 6-4(a). The scanning process is shown in Fig. 6-4(b), in which a camera video
Figure 6-4: (a) A human subject’s forearm covered with the temporary tattoo was scanned. (b) A camera video and an ultrasound video were recorded synchronously during scanning.

Figure 6-5: (a) Reconstructed skin surface and camera poses. The blue cameras correspond to keyframes, and the white camera is the one currently tracked. (b) The ultrasound images spatially registered in 6 DoF based on the camera pose estimates. (unit: mm)

and an ultrasound video were recorded at the same time during scanning. Here ultrasound gel was applied evenly on the skin surface as the ultrasound coupling medium.
Figure 6-6: (a) One of the ultrasound images showing the segmented artery. (b) Volume reconstruction of the brachial artery. The ultrasound scan trajectory is also shown.

Using the camera video, the algorithm reconstructed the skin map and the camera pose corresponding to each video frame, as shown in Fig. 6-5(a). Through ultrasound calibration (Section 4.2), pose estimates of the camera were converted to those of all the images in the ultrasound video, as shown in Fig. 6-5(b).

The brachial artery in the registered ultrasound images was segmented in Stradwin [117] by manually drawing contours around the artery. One of the ultrasound images with the contour is shown in Fig. 6-6(a). The surface-rendered artery is shown in Fig. 6-6(b) along with the ultrasound scan trajectory.

Similar experiments were performed on the thigh of a human subject with an additional layer of transparent dressing (3M Nexcare Tegaderm), as shown in Fig. 6-7(a). Ultrasound scans were conducted in freehand on the thigh to scan the femoral artery, following a roughly linear scan path. The reconstructed skin map and 6-DoF image poses are shown in Fig. 6-7(b). The femoral artery in the ultrasound images was manually segmented and surface-rendered in Stradwin.
Figure 6-7: (a) The scan region with artificial skin features in the in-vivo experiments. (b) Estimated skin feature points and camera trajectory (shown in blue) from scanning the region in (a). The corresponding ultrasound scan trajectory is also shown (in gray). (c) One of the recorded ultrasound images. The femoral artery (enclosed by the yellow rectangle) was manually segmented and surface-rendered. The reconstructed artery is shown in the inset.

One of the ultrasound images showing the artery is shown in Fig. 6-7(c), along with the reconstructed 3D volume of the artery.

It can be seen that in both the experiments, there are no significant irregularities on the surface of the reconstructed artery. Nevertheless, without the ground truth, it is hard to tell whether the smooth curvature is the true artery shape or an artifact resulted from estimation errors and probe compression. A methodology for qualitatively evaluating reconstructed volumes by comparison with real ultrasound images is presented in Section 6.3.

6.3 General Skin Surface with Natural Skin Features

6.3.1 Experimental Setup

The visual SLAM algorithm (Section 2.2) applied to natural skin features was validated by freehand scanning on three body parts of a human subject, including the lower leg, abdomen, and neck, which are illustrated in Fig. 6-8 along with the respective scan paths. Reconstruction of these scans is summarized in Table 6.1. It can be seen that, at the minimum, there were tens of inlier features in each frame.
Figure 6-8: Body parts where freehand scanning experiments were performed and the scan paths (shown in blue): (a) lower leg, (b) abdomen, and (c) neck.

for pose estimation, and the average inlier ratio was around 0.35, which requires less than 500 random samples in RANSAC to achieve a probability of 99.9%, for instance, of sampling four correct point correspondences at least once as required by the EPnP algorithm [63].

As the ground truth for the motion estimates from the system, independent measurement of probe motion during these freehand scans was obtained using OptiTrack V120:Trio (NaturalPoint Inc., Corvallis, OR, USA), a tri-camera 6-DoF tracking device with sub-millimeter accuracy. V120:Trio tracks a rigid body with five passive markers mounted to the probe, and as an example, the experimental setup on the lower leg is illustrated in Fig. 6-9(a). The homogeneous transformation matrix from the ultrasound image coordinate system \( U \) to that of the rigid body \( R \), denoted by \( T_{U_i}^R \), was determined in Stradwin [117] using the single-wall method [89] and remained the same throughout the scanning processes.

The tracked ultrasound image poses from the proposed system and V120:Trio were compared in the coordinate system of the first ultrasound image \( (i = 1) \), denoted by \( U_1 \), in order to examine effects of including ultrasound images in pose estimation (Section 2.3). To this end, the transformation from the \( i \)-th ultrasound image coordinates to \( U_1 \), denoted by \( T_{U_i}^{U_1} \), was computed at each time instance \( i \) using estimates from both the system \( (T_{U_i}^{U_1}(S)) \) and V120:Trio \((T_{U_i}^{U_1}(V))\). Their translational and rotational components were then compared. Note that since
Figure 6-9: (a) Setup for linear scanning experiments on the lower leg of a human subject. Probe motion was independently measured by OptiTrack V120:Trio as the ground truth. Involved coordinate transformations (green) and the coordinate system U1 (black) are illustrated. (b) Trajectories of the ultrasound image origin in U1 (in mm): ground truth (black), prior+camera+ultrasound (blue), and prior+camera (red).

The proposed system localizes the probe with respect to the skin surface, which is a reference coordinate system decoupled with the world coordinate system of V120:Trio, the body parts were immobilized throughout each scan to maintain the spatial consistency and allow for this comparison.

\[ T_{U1}^{U1(s)} \] can be computed by decomposing it into the following form:

\[ T_{U1}^{U1(s)} = T_{U1}^{U1} T_{C1}^{C1} T_{Ri}^{C1} \]  \hspace{1cm} (6.1)

\( T_{C1}^{C1} \) denotes the relative camera motion between the i-th and the first frame, which is estimated by the localization system. \( T_{Ri}^{C1} = T_{U1}^{C1} = (T_{C1}^{U1})^{-1} \) is estimated from ultrasound calibration and remains the same throughout the scanning process.

\( T_{U1}^{U1(c)} \) can be decomposed into the form:

\[ T_{U1}^{U1(c)} = T_{W1}^{U1} T_{R1}^{W1} T_{U1}^{R1} = (T_{R1}^{U1} T_{W1}^{R1})^{-1} T_{W1}^{R1} T_{U1}^{R1} \]  \hspace{1cm} (6.2)

\( T_{R1}^{U1} = T_{U1}^{R1} = T_{U1}^{R} \) is estimated from ultrasound calibration and remains the same throughout the scanning process. \( T_{R1}^{W1} \) denotes motion of R estimated by V120:Trio in its world coordinate system W at time instance i.
6.3.2 Motion Errors

Five freehand scans were performed on each one of the three body parts and examples of the pose estimation errors from the system are shown in Fig. 6-10. The drifting effect due to accumulated errors is observable, which is common for dead-reckoning approaches and is a shortcoming of the proposed system compared to optical or electromagnetic tracking. On average, for every probe travel distance of 10 mm, the accumulated drifting error is 0.91 ± 0.49 mm in translation and 0.55° ± 0.17° in rotation. A major cause of the estimation errors is the well-known ambiguity in discriminating between translation and rotation from camera frames, especially when there is a small camera FOV or insufficient depth variation in the scene [2]. Other main sources of errors include inaccuracies in ultrasound calibration (for estimation of $T_C^U$ and $T_R^U$) and body motion during scanning, such as motion caused by tremor, respiration, and heartbeat, especially for scans on the abdomen and neck.

In order to examine the effects of including ultrasound image regularity in pose estimation, as described in Section 2.3, probe pose estimation was performed twice in the system using the same data acquired in the lower-leg experiment: one with all the information (prior+camera+ultrasound), and one with only the prior motion model and camera frames (prior+camera). The translational components of $T_{UI}^{(S)}$ from both prior+camera+ultrasound and prior+camera, as well as those of $T_{UI}^{(V)}$ from V120:Trio, are shown in Fig. 6-9(b), which are equivalent to posi-
Figure 6-11: Estimation errors in probe motion versus the total probe travel distance (in mm) in the lower-leg experiment: prior+camera+ultrasound (blue) and prior+camera (red, dashed). (a) Translational errors along the X, Y, and Z axes (in mm) in coordinate system U1. (b) Rotational errors around the three axes (in degrees).

The pose estimation errors from the system, for both prior+camera+ultrasound and prior+camera, are shown in Fig. 6-11, where the rotational components are represented by Euler angles in the Z-Y-X system.

From Fig. 6-9(b) and Fig. 6-11, it can be seen that generally probe tracking was more accurate when ultrasound images were included in pose estimation. Specifically, with ultrasound images, the maximum translational and rotational errors are around 5 mm and 5 degrees, respectively, over a probe travel distance of 110 mm, as opposed to 10 mm and 10 degrees without the images. This observation is expected since ultrasound images provide additional constraints for pose estimation in all 6 DoF except Z-translation (U1 coordinates), which is eliminated in projection-based searches for corresponding pixel intensities between two ultrasound images (as described in Ultrasound Image Regularity in Section 2.3.) Specifically, improvement of tracking is significant in Y-translation, X-rotation, and Z-rotation, but not obvious in X-translation and Y-rotation. This behavior is mainly due to the fact that the spatial resolution of ultrasound images is much higher.
better along the Y-axis (axial) than along the X-axis (lateral). Additionally, tissue structures are usually shown as horizontal layers in ultrasound images, so motion along the Y-axis is more easily detectable than the X-axis. This information on probe motion along the Y-axis also helps improve the well-known inaccuracy in motion estimation along the optical axis from camera images.

### 6.3.3 Skin Feature Retrieval for Probe Relocalization

Ultrasound probe relocalization using the established skin map points as body landmarks highly relies on the distinctiveness of skin features and the performance was examined on the freehand scans. For the scan on each body part, the pose $T_{ul_i}^{(1)}$ was re-estimated at each time instance $i$ based on correspondences between skin feature points in the $i$-th frame and the set of map points, following the procedure described in Section 2.4. The absolute translational and rotational differences between this new pose estimate and the original estimate $T_{ul_i}^{(1)(S)}$ were then averaged over all the frames in the scan. The precision and recall for retrieving feature correspondences in the $i$-th frame were computed as follows and also averaged over all the frames:

\[
\text{precision} = \frac{\# \text{ inlier correspondences}}{\# \text{ all retrieved correspondences}} \quad (6.3)
\]

\[
\text{recall} = \frac{\# \text{ inlier correspondences}}{\# \text{ all map points visible in the frame}} \quad (6.4)
\]

These results are summarized in Table 6.2. Note that the keyframes were excluded from the computation since the map point descriptors are composed of feature descriptors extracted from the keyframes, and as a result, the recalls tend to be particularly high. The precision and recall from an example scan are shown in Fig. 6-12, where the spikes in recall at keyframes could be observed.

As can be seen in Table 6.2, the current operating point gave a high precision and low recall. Nevertheless, given the typically large number of map points visible in a frame (on the order of a thousand), this recall rate gives at least tens of
### Table 6.1: Summary of 3D Reconstructions for Freehand Scans

<table>
<thead>
<tr>
<th>Scan</th>
<th>Frames</th>
<th>Travel Distance (mm)</th>
<th># Keyframes</th>
<th># Map Points</th>
<th>Inlier Ratio (mm)</th>
<th>Inlier Ratio (degree)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lower Leg</td>
<td>430</td>
<td>111.62</td>
<td>25</td>
<td>4369</td>
<td>0.359 + 0.084</td>
<td></td>
</tr>
<tr>
<td>Abdomen</td>
<td>365</td>
<td>106.12</td>
<td>27</td>
<td>3279</td>
<td>0.369 + 0.086</td>
<td></td>
</tr>
<tr>
<td>Neck</td>
<td>146</td>
<td>37.92</td>
<td>8</td>
<td>2372</td>
<td>0.425 + 0.087</td>
<td></td>
</tr>
</tbody>
</table>

### Table 6.2: Results of Skin Feature Retrieval and Probe Relocalization for Freehand Scans

<table>
<thead>
<tr>
<th>Scan</th>
<th>Frames</th>
<th># Inliers/Frame</th>
<th>Precision</th>
<th>Recall</th>
<th>Trans. Error (mm)</th>
<th>Rot. Error (degree)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lower Leg</td>
<td>405</td>
<td>62.1 + 22.5</td>
<td>0.857 + 0.061</td>
<td>0.060 + 0.017</td>
<td>0.378 + 0.351</td>
<td>0.578 + 0.507</td>
</tr>
<tr>
<td>Abdomen</td>
<td>338</td>
<td>46.0 + 18.4</td>
<td>0.889 + 0.058</td>
<td>0.053 + 0.021</td>
<td>0.386 + 0.314</td>
<td>0.600 + 0.397</td>
</tr>
<tr>
<td>Neck</td>
<td>138</td>
<td>155.7 + 44.5</td>
<td>0.913 + 0.038</td>
<td>0.111 + 0.026</td>
<td>0.996 + 0.14</td>
<td>0.578 + 0.507</td>
</tr>
</tbody>
</table>
correct feature correspondences, which are sufficient for robust probe relocalization. Note that the recall for the neck scan is significantly higher than the other two scans, which is due to the shorter scan distance and thus less map point descriptors. To the descriptors belonging to a true feature correspondence, a smaller database of map point descriptors makes it less likely that there exists another sufficiently similar map point descriptor as defined by the inequality (2.19). Therefore, a correct correspondence is more likely to pass the filtering process based on descriptor similarity and be accepted as an inlier for relocalization.

### 6.3.4 3D Volume Reconstructions and Reslices

By using results from the localization system, 3D ultrasound volumes corresponding to the freehand scans were reconstructed. These 3D volumes were visualized using Stradwin [117], which also corrected probe pressure and jitter in motion estimation based on the non-rigid image registration method described in [118]. The effect of this correction on an example reslice from the lower leg scan is shown in Fig. 6-13.

Visualization and validation of the 3D reconstructions are shown in Fig. 6-14. On the left of Fig. 6-14 are the 3D ultrasound volume for each freehand scan (after correction of probe pressure and jitter) and a reslice plane roughly perpendicular
Figure 6-13: Example reslice from the lower leg scan (a) before and (b) after correction of probe pressure and jitter in motion estimation.

to the ultrasound image planes with the transducer at the top. The synthesized reslice is shown in the middle. 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 ultrasound images acquired at approximately the same positions and orientations as the reslice planes, which are shown on the right of Fig. 6-14. It can be seen that the tissue structures are consistent between the real ultrasound images and the portion of reslices highlighted by yellow. Although the required level of detail reproduction depends on specific clinical applications, this consistency shows the potential practical value of the proposed system in improving clinical workflows and aiding diagnosis by enabling the creation of diagnostically useful reslices after ultrasound scanning.

6.4 Summary

This chapter describes experimental results for freehand 3D US on human bodies. Motion tracking errors are quantitatively evaluated, which show that natural skin
features could be robustly tracked for probe localization. The quantitative evaluation of skin feature retrieval shows the distinctiveness of natural skin features and the potential to perform probe relocalization with respect to a previously acquired volume. Finally, the quality of reconstructed 3D volumes is qualitatively evaluated by comparing synthesized reslices with real ultrasound images. The consistency in tissue structures between the reslices and ultrasound images shows the system's ability to produce reslices at arbitrary angles that are clinically useful, which could potentially reduce the number of patient visits and improve clinical workflows.

Note that the results reported here are all from single-sweep scans, and it is possible to improve the drifting errors by scanning the same region multiple times. For instance, if after a single sweep, the probe is moved to a location where an ultrasound image has been previously acquired, one could determine the accumulated error so far, based on which the existing results could be adjusted. This technique is often referred to as “loop closure” and is usually composed of two parts: detection of a visited scene and backward correction. Interested readers are referred to [124, 7, 97, 110] for more details.
Figure 6-14: Visualization and validation of the freehand scan on (a) lower leg, (b) abdomen, and (c) neck of a human subject. Left: reconstructed 3D volumes and reslice planes. Middle: synthesized reslices. The red lines correspond to the ultrasound images acquired at time instance $i = 1$. Right: real ultrasound images acquired at approximately the same positions and orientations as the reslice planes. Note that the tissue structures are consistent between the real ultrasound images and the portion of reslices highlighted by yellow.
Chapter 7

Computer-Guided Ultrasound Probe Realignment

In addition to freehand 3D US, another application of the proposed probe localization system is computer-guided ultrasound probe realignment, which helps sonographers return the ultrasound probe to a position and orientation where an ultrasound image has been previously acquired. This system continuously localizes the ultrasound probe by tracking artificial skin features (described in Section 3.1) with the probe-mounted camera described in Chapter 4. The system then provides visual feedback indicating the deviation between the current and the target probe poses, which guides the sonographer to realign the probe.

The methods used in this realignment system are similar in concept to those in the freehand 3D US system. The implementations, however, are quite different since the realignment system requires real-time performance for providing visual feedback, whereas only off-line processing is performed in freehand 3D US. These design details, considerations, and the in-vivo evaluation are presented in this chapter. Part of this chapter has been published in [106].
7.1 Two-Frame Initialization

The realignment process starts by estimating the relative pose between the starting and the target camera poses. It is assumed that the starting pose is close to the target, so that there is overlap between the two FOVs. Feature correspondences are established between the starting and the target camera frames by using the SIFT method [65]. Radial distortion is corrected here and in later tracking by using the camera intrinsic parameters from standard camera calibration. From these correspondences, the relative camera pose and 3D positions of the feature points are estimated, up to a scaling factor, using the five-point algorithm [78] and triangulation [41], as shown in Fig. 7-1(a). Unlike in freehand 3D US, here the scaling factor for camera translation and 3D feature points is not calibrated (e.g. by using patterns with known dimensions as described in Section 2.2.5) since this value is not essential for probe realignment and is consistent throughout the process.

Estimates of the relative pose and 3D feature points are then iteratively refined by sparse bundle adjustment [64], in which the sum of re-projection errors $E_{\text{reproj}}$ is minimized:

$$E_{\text{reproj}} = \sum_{i=1}^{2} \sum_{k} D(x_{ik}, P(X_k, R_i, t_i)),$$

(7.1)

where $i$ indexes the camera frames and $k$ indexes the map points. $R_i$ and $t_i$ denote the rotation and translation of the camera, respectively, at the time frame $i$ is taken. $X_k$ is the $k$-th map point and $P(X_k, R_i, t_i)$ represents its perspective projection onto frame $i$. The function $D$ computes the squared Euclidean distance between this projection and the corresponding feature point $x_{ik}$ in frame $i$.

As the realignment process is initialized based on epipolar geometry, the performance generally improves with the baseline distance between the starting and target camera poses, provided that their FOVs sufficiently overlap. In fact, in the current system implementation, it is assumed that there is a non-zero baseline distance. If the two camera centers happen to be coincident (i.e. when there is no translational deviation between the two poses), the epipolar geometry is un-
defined and the system would fail to initialize the realignment process. Although in this case, the two poses could be related by a homography matrix instead of an essential matrix as in the five-point algorithm, criteria need to be designed to select one model between the two only based on the images [109].

7.2 Camera Pose Tracking

The feature points identified in the initialization stage are tracked in subsequent camera frames in the realignment process using the KLT (Kanade-Lucas-Tomasi) tracker [17], which then gives correspondences between the 3D map points and the 2D feature points in each camera frame. With these 2D-to-3D correspondences, the EPnP algorithm is applied to efficiently estimate the current camera pose relative to the target pose [63]. Here a maximum of 500 feature points are tracked for computational efficiency. The average re-projection error is typically around 1 pixel.

It should be pointed out that use of the KLT tracker, which is substantially more efficient than SIFT feature tracking as described in Section 2.2, is allowed by the application of artificial skin features, which provide a high-contrast and rich texture for feature tracking. In the case of tracking natural skin features, the KLT tracker will become much less robust since the tracked features are not obvious and thus the tracker performance will be highly susceptible to image noise. Although the CLAHE algorithm described in Section 3.2 could enhance the images of skin features, the resulting images would violate the brightness consistency assumption in the KLT tracker. Therefore, for tracking natural skin features, SIFT feature tracking on contrast-enhanced images was found to give better performance although it might require careful design and a fair amount of engineering to implement this method in a real-time system.

Feature points that fail to be tracked, possibly due to scene specularity, motion blur, or simply going out of the FOV, are discarded. It is assumed that the whole realignment process happens between the two initial poses (starting and target).
and therefore the number of feature points visible in the camera frame is always sufficient for pose estimation (at least four). This assumption holds as long as the probe is continuously moved toward the target pose in the realignment process.

### 7.3 Visual Guidance

In order to provide sonographers with intuitive visual feedback that indicates the current probe pose relative to the target, a static virtual pyramid is created amid the map of 3D points established in the initialization stage and projected onto camera frames, as shown in Fig. 7-1(a). In the user interface showing the current camera frame, the virtual pyramid would appear to be fixed with respect to the skin surface, and therefore its projection would change with the estimated camera pose.

For probe realignment, the sonographer could freely move the probe until this pyramid projection appears to be exactly aligned with that at the target probe pose, which indicates successful probe realignment. To assist this alignment process, the pyramid projection at the target pose is also shown in the current cam-
era frame. As shown in Fig. 7-1(a), edges of the virtual pyramid are drawn in different colors (red, green, and blue) to avoid ambiguity in aligning the pyramid projections. The pose and dimensions of the virtual pyramid are determined automatically based on the spatial distribution of 3D feature points. For instance, in this implementation, the pyramid is placed at the center of the point cloud and oriented with respect to the target pose. The pyramid base width is approximately half of the range spanned by the cloud. Note that theoretically, registration of three points is sufficient to determine a unique alignment, but here five points (namely, the five vertices of the pyramid) are used to provide more constraints for better performance.

Although the probe can be moved freely to align the pyramid projections, in practice, it was found that it was easiest to first correct the probe rotation (around the axial axis), then adjust the projection size by tilting the probe, and finally shift the probe to correct the translational deviation. As an example, a series of moves to realign the probe in Fig. 7-1(a) is illustrated in Fig. 7-1(b). Here the current pyramid projection is drawn in deep colors, and the target in light colors. When the two pyramid projections are aligned (i.e. successful probe realignment), the current projection will be changed to light colors and the ultrasound image at this pose will be automatically acquired. Here the pyramid projections are considered aligned when the Euclidean distances between corresponding pyramid vertices are all less than 3 pixels. This threshold value could be adjusted to balance the speed of performing probe realignment and the accuracy.

7.4 Implementation

In this system, the same hardware as described in Chapter 4 is used, including the ultrasound imaging system, camera, and probe housing. The algorithms were implemented in C/C++. Ultrasound image acquisition was automated using the Terason t3000 device with the Terason Software Development Kit (SDK). Camera frames and ultrasound images were acquired synchronously during realignment.
7.5 *In-Vivo* Experimental Results

The realignment system was tested on the thigh of a human subject. Artificial skin features were created by affixing a transparent dressing (3M Nexcare Tegaderm) to the skin surface, which was then covered by a temporary tattoo sticker with rich features. A high-contrast binary random pattern was used as the features, an example of which is shown in Fig. 7-2(a). The scan region with the artificial skin features is shown in Fig. 7-2(b). Standard ultrasound gel was applied evenly as the coupling medium.

At the start of the experiments, a reference camera frame and ultrasound image were acquired on the thigh, and this probe pose was considered the target pose. The freehand realignment process then started at a probe pose that was deviated from the target pose. There were no pre-determined constraints between the target and the starting poses other than the assumption that their FOVs overlap.

The camera-ultrasound image pairs acquired at the starting pose and the final realigned pose are shown in Fig. 7-3(a) and (b), respectively. The ground truth acquired at the target pose is shown in Fig. 7-3(c) for comparison. From the camera
Figure 7-3: The camera frames (top) and ultrasound images (bottom) acquired at the (a) starting, (b) realigned, and (c) target (ground truth) probe pose. The femoral artery is highlighted by yellow rectangles.

Frame in Fig. 7-3(a), it can be seen that at the start, the probe deviated from the target pose by a rightward translation and some counter-clockwise rotation, which resulted in the ultrasound image inconsistencies between the starting (Fig. 7-3(a)) and the target poses (Fig. 7-3(c)). (Note the horizontal shift of the femoral artery in the ultrasound images, for instance.) These inconsistencies were significantly reduced after realignment, as can be seen from comparison between Fig. 7-3(b) and (c).

The performance of probe realignment could be more easily observed by examining the absolute difference images of the ultrasound images in Fig. 7-3. Fig. 7-4(a) shows the difference image between the starting and the target ultrasound images (Fig. 7-3(a) and (c)). Darker shades indicate lower image difference while brighter shades indicate higher image difference; thus, a completely black difference image would indicate perfect agreement between the ultrasound images. The inconsistencies in the ultrasound images before realignment are clearly
Figure 7-4: (a) The difference image between the starting and target ultrasound images (Fig. 7-3(a) and (c)). The regions where the target image is brighter are shown in green, and those darker are shown in red. (b) The difference image between the realigned and target ultrasound images (Fig. 7-3(b) and (c)).

observable by the presence of brighter pixels, which are mostly resulted from misalignment of muscle fibers and the femoral artery. A significant improvement in alignment can be observed in Fig. 7-4(b), which shows the difference image between the target and the final realigned ultrasound images (Fig. 7-3(b) and (c)).

From the experiments, it was found that in this method, realigning translation along the camera optical axis (i.e. probe compression) is not as accurate as along other axes. This behavior is due to the fact that, in the perspective projection camera model, the pyramid projection is less sensitive to translation along the optical axis. Therefore, when judging successful realignment by a single threshold on vertex distances, a larger range of deviation in the compression level is tolerated. It is possible to mitigate this directional inconsistency in accuracy by setting intelligent criteria for successful realignment instead of a simple distance threshold.
7.6 Summary

This chapter describes application of the probe localization system to the task of computer-guided probe realignment, which involves guiding the sonographer to return the ultrasound probe to a position and orientation at which ultrasound images are previously acquired. The visual SLAM technique described in Section 2.2 is employed, and an intuitive user interface featuring a static virtual pyramid was designed to provide the required visual guidance. It is shown that after probe realignment, the tissue structures in the ultrasound image align better with the image acquired at the target pose.

Note that in the presented experiments, artificial skin features were applied to the scan region to provide high-contrast features, which allow real-time feature tracking. This practice is useful for image-based surgical guidance, where probe realignment is performed at most hours after the surgical intervention. The artificial skin features, however, last for less than one week and thus are not suitable for longitudinal studies, for instance. The development of ultrasound probe realignment based on natural skin features is part of the future work and is discussed in depth in Section 8.2.4.
Chapter 8

Conclusion

Ultrasound probe localization is essential for volume imaging, image-guided intervention, and longitudinal studies. The existing methods for probe localization, however, require expensive and bulky equipment. In addition, the probe motion is typically measured with respect to a coordinate system independent of the patient’s body, which makes it difficult to relocalize or realign the probe between imaging sessions. In this thesis, a highly cost-effective and miniature-mobile system for ultrasound probe localization has been presented, which determines the probe motion with respect to the skin surface. Hence, patient motion artifact is avoided and probe relocalization is allowed. Through a series of in-vitro and in-vivo experiments, performance of the system for freehand 3D US and probe realignment has been validated. Overall, it is shown that the proposed system could potentially be an alternative to conventional optical or electromagnetic tracking devices with a much lower cost and higher portability. It is therefore envisioned that this system could facilitate the use of freehand 3D US and longitudinal studies for early detection of diseases and quantification of disease progression.

8.1 Contributions

The contributions of this thesis have been described in the related chapters and are summarized in the following.
• **Skin Features as Body Landmarks**  This thesis proposes and validates the concept of extracting skin features as body landmarks. Design of artificial skin features is presented (Section 3.1), and the application of contrast enhancement techniques to images of skin surfaces for robust extraction of natural skin features is also demonstrated (Section 3.2). Through *in-vivo* experiments on human subjects, extraction of natural skin features with different skin tones, under varying exposure levels, and in multiple color spaces is further analyzed (Section 3.2.1 and 3.2.2). The distinctiveness of skin features in the context of probe relocalization is also examined (Section 6.3.3).

• **A Novel System for Freehand 3D Ultrasound**  The use of skin features allows the development of a highly cost-effective and miniature-mobile 6-DoF ultrasound probe localization system (Chapter 2), and the motion estimation errors have been evaluated on the human body (Section 6.3). Since this skin-based localization system determines probe motion with respect to the patient's body, it enables freehand 3D US that is robust to rigid patient motion. Volume reconstructions from this 3D US system have been validated on both phantoms and different human body parts, by using both artificial and contrast-enhanced natural skin features (Chapter 5 and 6). The synthesized reslices from the volume reconstructions have been shown to be structurally consistent with real ultrasound images, which suggests potential of the 3D US system in improving clinical workflows and aiding diagnosis (Section 6.3.4).

• **Body Landmarks for Probe Relocalization and Realignment**  The simultaneous localization and mapping framework for freehand 3D US further allows probe relocalization based on the skin map accompanying the acquired ultrasound volume (Section 2.4 and 6.3.3). In conjunction with the probe localization system, a real-time system that features an intuitive user interface with a virtual target has been developed to guide sonographers for probe realignment based on the established skin map (Chapter 7).
• **Framework for Information Fusion** For optimal probe pose estimation, a Bayesian probabilistic framework has been developed to incorporate information from camera images of skin features, ultrasound images, and a prior motion model (Section 2.3). Through experiments on the human body, it has been shown that this information fusion improves motion estimation error in both translation and rotation (Section 6.3).

• **Camera-Probe Hardware Design** System prototypes for both freehand 3D US and ultrasound probe realignment have been developed. A probe housing was designed and two novel calibration methods were developed (Chapter 4 and Section 7.4). Through experiments on the human body, it was shown that this hardware design avoids skin surface deformation due to probe contact in the camera FOV and hence allows reconstruction of an undeformed skin map.

### 8.2 Limitations and Future Work

In this section, limitations of the current prototype systems are summarized, and some potential improvements are discussed as the future work. Other design possibilities have also been discussed in the summary section of each corresponding chapter.

#### 8.2.1 Defocus Blur

Cameras have limited depths of focus, and hence in the current hardware configuration as shown in Fig. 4-1, defocus blur can become severe when the ultrasound probe is tilted beyond a certain range of angles (typically from -20 degree to +20 degree). It is possible to use an auto-focus camera, instead of a fixed-focus one, to address this defocus issue. The research effort can be divided into two parts: auto focus during video acquisition and visual SLAM with variable focal lengths.
Auto Focus during Video Acquisition

Algorithms that detect occurrence of defocus blur and automatically adjust the focal length to maintain focus at all times need to be developed and integrated into the video acquisition software. Detection of defocus blur could potentially be achieved by analyzing the frequency content of the camera FOV. The algorithm also needs to decide the direction and magnitude in which the focal length needs to be changed, and potentially to integrate motion predictive models to perform focus adjustment efficiently. Hardware and software co-design could be explored to optimize the performance.

Visual SLAM with Variable Focal Lengths

In the future work, the current visual SLAM algorithm could potentially be extended to video frames with varying focal lengths, as acquired by the auto-focus system described above. Although the majority of existing visual SLAM algorithms assumes a fixed focal length during video acquisition [25, 56, 77], it has been shown that monocular visual SLAM is possible with varying, and even unknown, focal lengths [100, 128].

8.2.2 Fusion of Multiple Information Sources

Although a single camera has been sufficient for complete 6-DoF probe localization, the reconstruction accuracy and robustness could potentially be improved by appropriately incorporating other information sources. Incorporation of ultrasound images and a prior motion model has been presented in Section 2.3. In the following, the future work on fusion of multiple cameras, inertial measurements and ultrasound speckle decorrelation is described.

Multiple Cameras

Due to existence of noise in detected motion flows, there are inherent ambiguities in camera pose estimation purely from camera frames [2]. For instance, motion
flows due to camera rotation around one image axis are similar to those due to translation along the other axis. Therefore, given the corresponding field of motion flows, it is difficult to discriminate between the two kinds of motion accurately. Nevertheless, it is possible to fuse motion flows from multiple cameras to resolve this ambiguity, instead of using a single camera as in the current system. For instance, a second camera can be mounted to the ultrasound probe facing the skin surface, rotated by 90 degrees from the first camera (shown in Fig. 8-1), to provide additional information for resolving the aforementioned ambiguity. Similar research has been performed and the preliminary results were reported in [54].

Inertial Measurements

Inertial measurements are particularly promising for performance improvement of the proposed system, as it has been found that visual and inertial measurements often complement each other well in camera pose estimation [103, 38, 81]. Visual systems generally provide accurate and long-term reconstruction of camera motion and the scene, but tracking of features could be significantly affected by motion blur due to rapid motion. On the other hand, inertial sensors pro-
vide good signals during rapid motion but suffer from biased estimation due to accumulation of drift, especially in the translational component, where double integration of the acceleration reading is required. The standard framework for online estimation is to maintain a state vector that describes camera motion in an extended Kalman filter (EKF) incorporating both visual and inertial measurements. As visual sampling (around 30 frames/second) is often much slower than inertial sensing (around 100 samples/second), a filter bank is usually applied to perform multi-rate processing [38].

Inertial measurements could also be naturally incorporated into the Bayesian framework described in Section 2.3 as an additional energy term in the total energy function $E_{\text{total}}$. The related energy $E_{\text{inertial}}$ describes the distance between inertial measurements and the estimated probe motion. For instance, terms for rotation $E_R$, translation $E_t$, and linear velocity $E_i$ could be included:

$$E_{\text{inertial}} = E_R + \alpha E_t + \beta E_i$$

(8.1)\[E_R = D_R(R_i, R'_i(\omega(\tau)))\]

(8.2)\[E_t = D_t(t_i, t'_i(a(\tau), g))\]

(8.3)\[E_i = D_i(i_i, i'_i(a(\tau), g))\]

(8.4)

The variables $R_i$, $t_i$ and $i_i$ denote estimates over which optimization is performed. The measurements from inertial sensing, denoted by $R'_i$, $t'_i$ and $i'_i$, are computed from kinodynamic models using the measured angular velocity $\omega(\tau)$ and linear acceleration $a(\tau)$, where $\tau$ denotes the time. The vector $g$ denotes the gravity. The distances between these estimates and inertial measurements are computed by the respective distance functions, which are denoted by $D_R$, $D_t$, and $D_i$. The three terms are weighted by $\alpha$ and $\beta$.

**Ultrasound Speckle Decorrelation**

In addition to external sensors, ultrasound speckle patterns also provide measurements of the relative motion between ultrasound images [46]. The speckle
pattern in an ultrasound image patch is a function of the scatterers in the resolution cell around the corresponding region. As the spatial resolution of ultrasound imaging is limited by beam focusing, neighboring ultrasound patches often have overlapping resolution cells. Depending on the overlapping between cells, the speckle patterns in those patches are partially correlated. Therefore, by analyzing correlation of the speckle patterns in two neighboring ultrasound images, it is possible to obtain estimates of the relative motion.

The estimation of in-plane motion (axial and lateral translation, and rotation around the elevational axis) could be performed by standard speckle tracking between ultrasound images and the techniques have been well-studied [90]. Out-of-plane motion could be estimated by the speckle decorrelation method, which is based on the fact that the correlation between neighboring speckle patterns decreases with the elevational spacing [37]. Although currently the motion tracking performance given by the decorrelation method alone is not comparable to that from external motion sensors, incorporating a decorrelation term into the Bayesian framework described in Section 2.3 could potentially improve the overall motion estimation.

### 8.2.3 Discrete-Continuous Optimization

The visual SLAM algorithm described in Section 2.2 is often referred to as the incremental bundle adjustment approach and is widely used for structure-from-motion tasks [100]. However, this incremental approach could produce different results depending on the order in which the images are added and sometimes leads to reconstruction failure due to local minima in optimization. These observations lead to the development of a unified framework that involves performing large-scale global optimization at once, with the aim of improving the system robustness to reconstruction failure and allowing natural incorporation of additional constraints [23].

In this formulation, both the absolute 3D map point positions \(X_k\) and camera
poses ($R_i$ and $t_i$) with respect to a given coordinate system are hidden variables in a Markov random field (MRF), which are represented by nodes in the graph as illustrated in Fig. 8-2. The coordinates of the $k$-th map point in the $i$-th camera coordinate system are denoted by $X_{i,k}$ and can be expressed as:

$$X_{i,k} = R_i X_k + t_i.$$  \hfill (8.5)

The $i$-th and $j$-th camera poses are constrained by a cost function that characterizes the distance between their estimated and measured relative pose, which is represented by an edge $e_{ij}$ between the two corresponding nodes in the MRF. The estimated relative pose can be computed from the absolute pose estimates:

$$R_{ij} = R_j R_i^T$$ \hfill (8.6)

$$t_{ij} = t_j - R_j R_i^T t_i$$ \hfill (8.7)

Measurements of the relative pose could be obtained from both camera images and inertial sensors, for instance, which are denoted by $(R_{ij}^{\text{visual}}, t_{ij}^{\text{visual}})$ and $(R_{ij}^{\text{inertial}}, t_{ij}^{\text{inertial}})$, respectively. The rotational component of the total distance $D_C(R)$ and
the translational component $D^t_C(T)$ are then expressed as:

$$D^R_C(\mathcal{R}) = \sum_{e_{ij} \in E_C} \left[ d^R(R^\text{visual}_{ij}, R_i \dot{R}^T_j) + d^R(R^\text{inertial}_{ij}, R_i \dot{R}^T_j) \right] \quad (8.8)$$

$$D^t_C(T) = \sum_{e_{ij} \in E_C} \left[ d^t(t^\text{visual}_{ij}, \dot{t}_j - R_i \dot{R}^T_i t_i) + d^t(t^\text{inertial}_{ij}, \dot{t}_j - R_i \dot{R}^T_i t_i) \right] \quad (8.9)$$

where the subscript $C$ denotes the constraints between camera poses. $\mathcal{R}$ and $T$ are the sets of rotational and translational estimates, respectively. $E_C$ denotes the set of all the edges between the nodes for camera poses. The functions $d^R$ and $d^t$ compute the distances for the relative rotations and translations, respectively.

The total distance $D_C(\mathcal{R}, T)$ is then the combination of $D^R_C(\mathcal{R})$ and $D^t_C(T)$ with a weighting coefficient $\alpha_1$:

$$D_C(\mathcal{R}, T) = D^R_C(\mathcal{R}) + \alpha_1 D^t_C(T) \quad (8.10)$$

Note that in addition to camera frames and inertial sensors, constraints from ultrasound images and a prior motion model could also be included in this distance function, as described in Section 2.3.

The spatial configuration between the $i$-th camera pose and the map point $X_k$ is constrained by a cost function based on epipolar geometry, which is represented by an edge $e'_{ik}$ in the graph. Ideally the vector from the origin of the $i$-th camera coordinate system to $X_k$ (i.e. $X_k + R_i \dot{R}^T_i t_i$) is parallel to the ray from the origin to the image point $x_{ik}$, which can be expressed as $R_i^T K^{-1} x_{ik}$. ($K$ denotes the matrix of camera intrinsics.) Therefore, the cost can be defined as the angle between the two vectors, which is computed by the distance function $d^\theta$:

$$D_p(\mathcal{R}, T, \mathcal{X}) = \alpha_2 \sum_{e'_{ik} \in E_P} d^\theta(X_k + R_i \dot{R}^T_i t_i, R_i^T K^{-1} x_{ik}) \quad (8.11)$$

The subscript $P$ denotes the constraints between map points and camera poses, and $E_P$ denotes the set of all the edges between the two sets of nodes. $\mathcal{X}$ denotes the set of estimated 3D map points. $\alpha_2$ is a weighting coefficient.
In addition to the above binary constraints, prior information about the absolute camera orientations or known map points can be naturally incorporated into this MRF framework by adding unary constraints (denoted by the subscript \(U\)). For instance, the distance functions for orientation measurements \(\hat{R}_i\) from magnetometers and map point measurements \(\hat{X}_k\) from known fiducials could be written as:

\[
D^R_{U}(\mathcal{R}) = \alpha_3 \sum_i d^R(R_i, \hat{R}_i) \\
D^X_{U}(\mathcal{X}) = \alpha_4 \sum_k d^I(X_k, \hat{X}_k),
\]

where \(\alpha_3\) and \(\alpha_4\) are weighting coefficients.

Given this MRF formulation, the optimal camera poses and map points could be found through discrete-continuous optimization [62, 23]. Minimization of the total cost is initialized in a discrete manner by loopy belief propagation on the graph, which provides an initial guess for the subsequent continuous refinement through non-linear least squares minimization methods like the Levenberg-Marquardt algorithm. It has been shown that, compared to the incremental approach, this batch approach is potentially more robust to reconstruction failure and scales better as the number of images grows [23].

### 8.2.4 Probe Realignment based on Natural Skin Features

The current implementation of the ultrasound probe realignment system relies on artificial skin features, which are created by affixing a transparent dressing to the skin surface covered by a temporary tattoo sticker with high-contrast random patterns. The duration between the initial image acquisition and probe realignment could be no longer than several days, which is suitable for surgical intervention but poses a challenge to the clinical use of this system for longitudinal studies.

It is possible to improve the probe realignment system and facilitate its use in clinical settings by using contrast-enhanced natural skin features as body land-
marks, instead of artificial features. In Section 6.3.3, it has been shown that contrast-enhanced natural skin features are highly distinctive from each other, and therefore the ultrasound probe could be relocalized with respect to a skin map constructed during a freehand 3D US scan based on recorded skin features. It is envisioned that, by integrating the work on repeated recognition of natural skin features and the intuitive user interface, a probe realignment system could be developed that supports accurate longitudinal studies over a period of multiple years.

The major challenge in this research work is the real-time performance that is required for giving instant visual feedback. For off-line processing, it is possible to perform contrast enhancement on the images of natural skin features before feature detection and descriptor extraction. It is challenging, however, to perform this image enhancement in real time. One possible solution is to implement the system using GPU (graphics processing unit) programming, which parallelizes the computation. Another possibility is to use an optical or electromagnetic tracker for continuous and real-time probe tracking, where the initial estimation of relative probe motion using natural skin features serves as a calibra-
tion step. Note that in order to integrate measurement from an optical or electromagnetic tracker with skin-based probe localization, the scale factor in skin-based reconstruction needs to be determined, which could be achieved by capturing two current camera frames, as illustrated in Fig. 8-3.

The relative pose between the target and first current camera frame, denoted by $T_1$, serves as the calibration that allows representing the target pose in the coordinate system of the tracker. This relative pose could be estimated based on skin features but only up to a scaling factor. In order to address the scale ambiguity, a second current camera frame is acquired in the calibration step. Although based on the skin features, the relative poses between the three frames (i.e. target, first, and second) are all estimated up to a scaling factor, scale of the relative pose between the first and second current camera frames (denoted by $T_2$) is given by the tracker. Therefore, the scale of $T_1$ could be determined as well. Performance of this skin-based probe realignment approach could potentially be improved by registration of tissue structures in ultrasound images or volumes, similar to the approach described in Section 2.3.
Bibliography


