UNIVERSITY OF CALIFORNIA SAN DIEGO

A Vision-Based Approach to Scan Target Localization for Autonomous Lung Ultrasound Imaging

A thesis submitted in partial satisfaction of the requirements for the degree Master of Science

in

Electrical Engineering (Machine Learning and Data Science)

by

Jianzhi Long

Committee in charge:

Professor Truong Nguyen, Chair
Professor Cheolhong An
Professor Imanuel Lerman
Professor Xiaolong Wang

2022

The thesis of Jianzhi Long is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2022

DEDICATION

I dedicate my thesis to my family. An utmost gratitude to my parents, whose unwavering encouragement and expectation for excellence carried me through the toughest part of this journey. Even though one Pacific Ocean away, their support has always reached me instantly when needed.

I dedicate this thesis and give special thanks to my project partner and dear friend Benny, who shared with me a tremendous workload in the process, and inspired me to constantly strive for a better version of myself with his perseverance.

I also dedicate this work to the Rock-'N'-Roll legends, including AC/DC, Def Leppard, Guns N' Roses and Cui Jian, whose passionate tunes and heartfelt lyrics have shaped my attitude on work and life in a profoundly positive way during this period.

EPIGRAPH

It's a long way to the top
if you wanna rock 'n' roll.

AC/DC

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I wish to thank my committee members who were more than generous with their expertise and precious time. A special thanks to Professor Truong Nguyen, my committee chairman and supervisor, for kindly admitting me into the Summer Research Internship Program that has led to this thesis, and consistently providing constructive feedback on my work. Thank you Professor Imanuel Lerman for sharing knowledge and equipment on lung ultrasound imaging. Thank you Professor Cheolhong An and Professor Xiaolong Wang for agreeing to serve on my committee.

I would like to send my gratitude to all the coauthors, especially Mr. Jicang Cai for your countless hours spent on this work. Thank you Abdullah Al-Battal and Shiwei Jin for your insightful opinion and encouragement. Thank you Dr. Jing Zhang and Dr. Dacheng Tao for your help and advice on human pose estimation and the writing process. Thank you Professor Truong Nguyen for overseeing this research project.

Finally, I would like to acknowledge and thank the Electrical and Computer Engineering department of UCSD, for allowing me to conduct research and present it in the format of Master Thesis. A special thanks to Ms. Chudan Li, my MS Advisor, for helping me with the logistics.

This thesis is currently being prepared for submission for publication of the material. Long, Jianzhi; Cai, Jicang; Al-Battal, Abdullah; Jin, Shiwei; Zhang, Jing; Tao, Dacheng; Nguyen, Truong. “A Vision-Based Approach to Scan Target Localization for Autonomous Lung Ultrasound Imaging”. The thesis author was the primary investigator and author of this paper.

VITA

2020 Bachelor of Science in Electrical Engineering, University of Illinois at Urbana-Champaign

2022 Master of Science in Electrical Engineering (Machine Learning and Data Science), University of California San Diego

FIELDS OF STUDY

Major Field: Electrical and Computer Engineering (Machine Learning and Data Science)

ABSTRACT OF THE THESIS

A Vision-Based Approach to Scan Target Localization
for Autonomous Lung Ultrasound Imaging

by

Jianzhi Long

Master of Science in Electrical Engineering (Machine Learning and Data Science)

University of California San Diego, 2022

Professor Truong Nguyen, Chair

Ultrasound is progressing toward becoming an affordable and versatile solution to medical imaging. In recent years, the need for a fully autonomous ultrasound system has been accelerated due to its labor-intensive nature and the advent of COVID-19 global pandemic. In this work, we tackle the important yet seldom-studied problem of scan target localization, under the setting of lung ultrasound imaging. We propose a purely vision-based, data driven method that incorporates learning-based computer vision techniques. We combined a human pose estimation model with a specially designed interpolation model to predict the lung US scan targets, while multi-view stereo vision is deployed

to enhance the accuracy of 3D target localization. We collected data from 30 human subjects for testing, and obtained satisfactory result from test experiments, achieving a success rate above 80% for all scan targets under an error threshold of 25mm. Finally, our approach can serve as a general solution to other types of US scan, with many potential improvements in terms of model complexity and runtime. The code is available at this url.

Introduction

Ultrasound (US) imaging has been increasingly prevalent in modern day clinical practice. Nowadays, the scope of US image analysis has covered almost all locations on the body, including brain, thyroid, heart, breast, fetal and prostate [4].

Besides US, the other medical imaging technologies that doctors rely on are X-ray Projection Imaging, Computed Tomography (CT), Nuclear Imaging and Magnetic Resonance Imaging [14]. In comparison, US imaging is harmless, cost effective, portable and can provide real-time feedback. Because of these qualities, US has been deployed widely in hospitals, especially in the urgent care section and remote areas where it is too costly to afford the other alternatives [26]. Also, a significant amount of research has been done in the recent years to address its drawbacks, including high level of noise, visual artifacts and small detection radius [4]. Therefore, US shows great promise in the future of medical imaging.

So far, US imaging is mainly performed by trained sonographers and is very labor intensive [54]. The quality of ultrasound image is strongly dependent on the skill level of the human operator [52]. As a result, in the past two decades researchers have begun to explore the application of robotics in ultrasound applications [44]. Robotic manipulators have the potential of achieving accuracy, consistency, dexterity, and maneuverability simultaneously, which perfectly compensates the shortcomings of manual acquisition. However, many of these robotic systems either provide aid to freehand scanning, or require an operator to input instructions [48, 55, 35].

US has also been considered as the standard of care for reliably identifying lung

pathology indicative of respiratory infection and disease progression. While a significant cost is involved in training professional sonographers, there is also a real possibility for contagious diseases to propagate from patients to the medical staff. The need to eliminate the risk of infection between patients and medical staff becomes much more urgent with the advent of COVID-19 global pandemic. Since people with COVID-19 infection can be asymptomatic, all healthcare workers who undertake face-to-face clinical work are potentially at risk [32]. Therefore, in the age of a global pandemic, there is a pressing demand for a fully automatic US imaging system that could protect both the medical staff and patients.

Despite being specialized to different types of US imaging, some generality remains for most autonomous US imaging systems [26]. First, a scan-path planning algorithm defines a set of waypoints for US probe, based on data from sensors like RGB-D cameras. A position and force control module is in place for adjusting the US probe, to ensure optimal contact for acquiring high quality US images and the safety of the subject. Finally, US imaging processing techniques are applied to enhance image quality for further evaluation. There are many works that address position and force control of US by fine-tuning the position and orientation of US probe, usually through feedback from force sensors [1, 34, 21, 12]. Moreover, a substantial number of effort has been invested in the topic of US image processing, including image segmentation [20, 30, 58, 17], pathology detection and classification [47, 2, 8, 46], and 3D volume reconstruction [50, 33].

In comparison, there are very few works on automatic scan path planning. It typically consists of two steps: first, the transducer probe is moved to the proximity of the target scan location, then a scan path can be constructed in the neighboring region of the starting point. The first step, known as the scan target localization problem [31], is particularly crucial for achieving truly autonomous US scan and yet still very much underexplored. The transducer probe needs to be placed in the close proximity of the starting point of the scan in the first place in order to initiate the following scanning

process. Until now, there are only a few works that address this issue [18, 24, 59, 31]. The reason behind this lack of research effort might be the lack of relevant data in comparison to the wide range of US imaging tasks and the large variability in human anatomy among patients. So far, even though existing work has explored a variety of methods, none of them could be readily adopted as a general US imaging solution for human subjects.

In this work, we develop an accurate scan target localization system via a learning-based method with visual sensors only. Visual sensors are affordable and widely available, matching the advantages of US imaging itself. In addition, the advance and widespread application of computer vision technology provides a diversified sets of available tools. We assume that the image of human torso can provide rich information on the location of human internal organs. Even though humans as a species share a normality of anatomy to a certain degree, variability among individuals is still significant [63]. As a result, we hope to learn this relationship through a data driven approach and thus directly extract the target scan locations from the visual data.

We propose a scan target localization system for Lung Ultrasound Scan, which integrates the traditional CV techniques and the recent deep learning approaches. The system incorporates an ultrasound machine, two RGB-D cameras, a 6-DOF industrial robot, and a computer that controls the visual sensors and robot actuator. We consider US positioning as a variant to the existing human pose estimation problem and propose an algorithm that integrates a pose estimation model with an scan target interpolation model with empirically determined parameters. multi-view stereo vision is another crucial component in our system, as it recovers the 3D information of the features extracted from 2D images. The workflow of the system is as follows:

1. Two-view color and depth images of the body are captured.

2. The locations of human body keypoints in the color images are extracted by pose estimation.

3. 3D coordinate of these keypoints is computed via triangulation, and then set as the input of the scan target interpolation algorithm.

4. Depth data is used for 3D reconstruction, upon which a Nearest Neighbor method is employed for both normal vector estimation and refining 3D target computation result.

We tested our system on 30 human subjects with diverse body composition and optimized our model parameters with collected data, which improves the accuracy of the results. We also discuss the potential of our method as a general and readily applicable solution to all types of US scan. In summary, our contribution is three-fold and can be summarized as follows:

- We present a vision-based autonomous lung US scan target localization system, which could also be generalized to other types of US scans.

- We approach the scan target localization problem as a variant of the human pose estimation problem.

- We deploy multi-view stereo vision to enhance the accuracy of the target estimation.

Related Works

We consider our work as an interdisciplinary application of computer vision in the field of robotic medical image acquisition. Usually, computer vision techniques including matured image processing algorithms and the recent deep learning models are applied on the already acquired medical images, yet much less work has been done regarding the automation of the acquisition process itself. We would like to build a bridge that meets the growing demand of autonomous image acquisition with the ample availability of computer vision technology.

The core of our system is a target computation algorithm based on human pose estimation. We extend the model for traditional human pose estimation problem to locate US scan targets on human body. Also, our solution is intended to serve as an integral component of a standard, generalizable, fully autonomous US imaging system for all types of US scans.

Human pose estimation (HPE) is widely used in areas ranging from virtual reality to medical assistance. It aims to automatically locate the human body parts from images or videos [10]. In recent years, both 2D and 3D HPE methods have been developed with the aid of deep learning [7]. The outputs of 2D HPE models are the image pixel coordinates of the respective body part keypoints, whereas the outputs of 3D HPE models are the 3D coordinates of the keypoints. Compared to 2D HPE datasets, 3D HPE datasets are scarce and are usually obtained under constrained environments with limited generalizability. As a result, we use a state-of-the-art 2D HPE model combined with depth information directly obtained from multi-view stereo vision to compute the 3D coordinate of the target

locations, a setting for which we believe could attain maximum accuracy.

2D HPE algorithms consists of two main categories: the top-down approach and the bottom-up approach [7]. The top-down method consists of a human body region detector and a single person pose estimator [11, 42, 19, 51]. The detector outputs bounding boxes surrounding each of the human subject in the image. The pose estimator is then run through each cropped bounding box to obtain the corresponding keypoint locations. On the other hand, the main components of most bottom-up methods include body joint detection and joint candidate grouping [6, 38, 22, 23]. The algorithm first predicts all the 2D joints present in the image and then assembles them into independent skeletons. In comparison, the runtime of top-down methods is often much slower, and has a linear relationship to the number of people detected in the image. However, the top-down methods usually yield better results, setting state-of-the-art performance for most benchmarks.

The metrics used to evaluate the performance of HPE models mainly include Average Precision (AP), Average Recall (AR) [7]. A predicted joint location is considered as a true positive if it is within a threshold of the ground-truth [60]. AP and AR are then computed for each body joint after evaluating all the joint predictions. Some common datasets for training 2D HPE models are FLIC [49], MPII [3], COCO [27], AIC [56]. Recently, the publication of OCHuman [61] and CrowdPose [25] datasets has introduced more challenges in terms of body occlusions and presence of a crowd in the image.

The existing vision-based US scan target localization methods can be divided into two categories: non-learning based and learning based. For non-learning based methods, Huang et al. [18] proposes a general solution that has been tested on breast, lumbar, thyroid and fetus. A single fixed Kinect camera captures the color and depth data of the scene. Only phantom experiments are conducted and a rule-based image segmentation method based on K-means clustering is used to extract the contour of the phantom in the color image to locate the scan region. This method is not applicable to real human subjects as the human body is continuous and cannot be physically divided into separate

sections.

In constrast, Lee et al. [24] specifically target for breast US scan. Their method also uses a single Kinect camera, but requires the breast of the patient to be aligned to a specific detection region, where the scan region is then determined by features including nipple locations and skin colors. This approach cannot generalize to other body parts, and it is difficult for the fixed detection region to generalize to large variations in individual breast formations.

There exists three different approaches for learning-based scan target localization methods. Focusing on US scan for human spine, Yang et al. [59] develop an end-to-end deep learning based model that directly extracts the segmentation map of the entire longitudinal human spine area from the color and depth data. The scan path is generated from the segmentation map, and the 3D coordinate of scan points is then deprojected from their 2D pixel coordinate. To train the model, data collected from 10 human subjects has been augmented to form a training set of size 2500. This method can be applied to other parts of the human body such as knee joint and thyroid, however those areas generally have clear visual and depth features and are less deformable. As a result, it could be challenging to achieve accurate result on more spreadout and deformable areas on human body such as breast and abdomen.

A Reinforcement Learning (RL) based method was proposed by Ning et al. [39]. The state information of the environment includes camera image, US image and force measurement. With the US image and force measurement pre-encoded into RGB image space, the RL model takes only a single RGB image as input and outputs the action of the US probe, without the need to estimate the robot target pose. The model is able to guide the probe to the subject from distance and perform adjustment after making contact. Camera calibration and registration are not necessary, and fixed camera pose is not required. This method has the potential to become a general solution for US imaging in the future, provided that vast amount of multi-modal data from a diverse subject profile

is available and can be matched by a equally sophisticated model. However, so far all of the training data was collected with phantoms only and under strict background setting, therefore currently this approach has very limited generalization ability.

Lastly, Ma et al. [31] provides a HPE based approach for lung US, which utilizes a single-view RGB-D camera. A DensePose model [13] is deployed to map the human anatomical locations as pixels in the 2D image to vertices on the 3D Skinned Multi-Person Linear (SMPL) mesh model (UV-coordinates) [29]. By assuming that the UV-coordinate of the scan targets is universal for all individual subjects, fixed target values are directly encoded into the algorithm as ground truths. For each image of the subject, the pixels with the same UV-coordinate as the ground truth are selected and averaged to obtain the pixel coordinate of the scan target, which is then deprojected to 3D coordinate. The main problem of this work lies in the assumption of the author on the fixed ground truth. Besides this, all experiments are conducted on a single phantom, which further detracts the prospect of the general application of this method.

Methods

Target Scan Locations

Fig. 1 shows the target scan locations according to the 9-point pulmonary ultrasound protocol specified by Tierney et al.[53]. The nine locations include front and lateral points, and are spread across both the left and right side of the human body. Due to the limited reach of our robot arm, we will focus on the scan locations on the right side of the torso. Also, performing US scan on the Posterior Axillary Line (PAx) would be inconvenient for all patients, and infeasible with those relying on ventilators. As a result, we will be testing our method on scan locations 1, 2 and 4, situated on the Mid-Clavicular Line (MC) and Anterior Axillary Line (AAx).

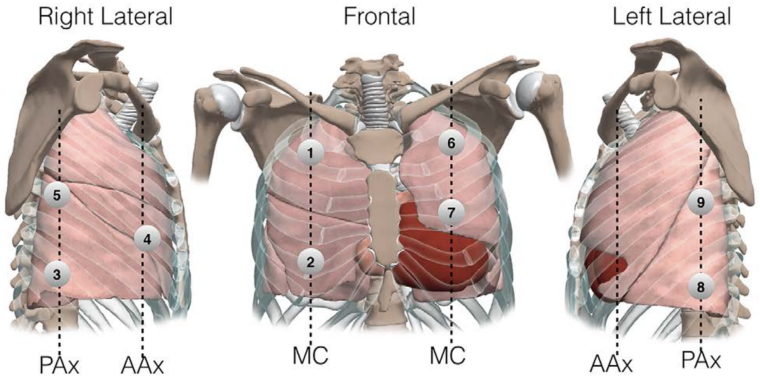


Figure 1. 9 Target Scan Locations. AAx = anterior axillary line, PAx = posterior axillary line, MC = mid-clavicular line (image adopted from [53]). We will focus on Target 1, 2, 4.

System Design

Our system consists of the following apparatuses: an industrial 6-DoF robot arm (Universal Robots UR3e), a Verasonics US machine paired with a Verasonics L12-3V linear array transducer, two Intel RealSense D-415 cameras, a PC (Dell Inspiron 16 Plus) that controls the visual sensors and robot actuator, a camera stand and a stretcher (see Fig. 2). The robot arm can handle a maximum payload of 3kg and has a maximum circular moving range of radius 500mm. We use a UR python library developed by Rope Robotics (Denmark) to program and manipulate the robot arm. The transducer probe is mounted on the robot arm gripper with a customized holder. The two depth cameras are mounted on the camera stand to capture images of the subject lying on the stretcher. The depth cameras have an ideal sensing range from 0.5m to 3m, and their resolution is set to 640 × 480 pixels. The target scan poses are computed from the captured two-view RGB-D images, to which the robot arm is programmed to move.

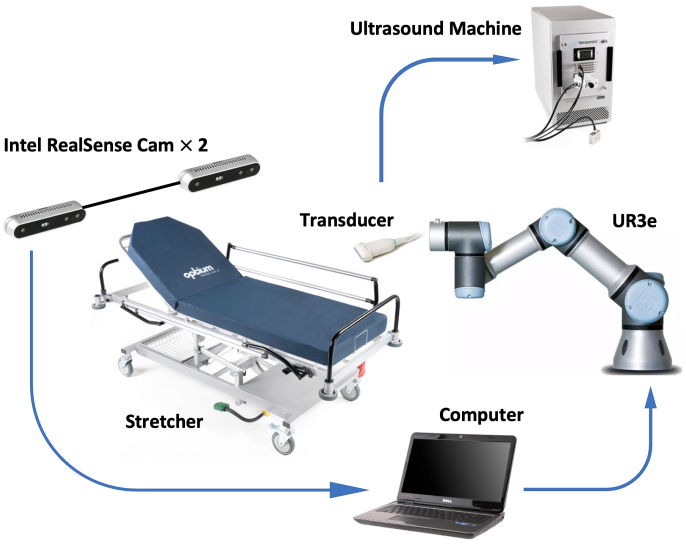


Figure 2. Apparatus setup.

Pipeline

Fig. 3 illustrates the general pipeline of our method. First, we perform hand-eye calibration to obtain the transformation between the camera coordinate frame and the robot base coordinate frame, and the latter is set as the global coordinate frame. Then, two-view color and depth images of the subject lying on the stretcher are captured from the two cameras. A pose estimation model is used to extract the 2D keypoint features of the subject. A 3D point cloud of the two-view RGB-D images is constructed along with the surface normal vector estimation of all the points. We leverage the 2D and 3D information to compute the coordinates of the target scan locations and its surface normal vector in the global frame. Specifically, triangulation algorithm is applied to compute the 3D global coordinates of keypoint features present in both color images, which are then used to interpolate the target scan locations. Aside from hand-eye calibration, the rest of the process is performed twice, for scan locations on the front (AAx) and side (MC) of the body, for which the subject is asked to maintain different postures during experiment.

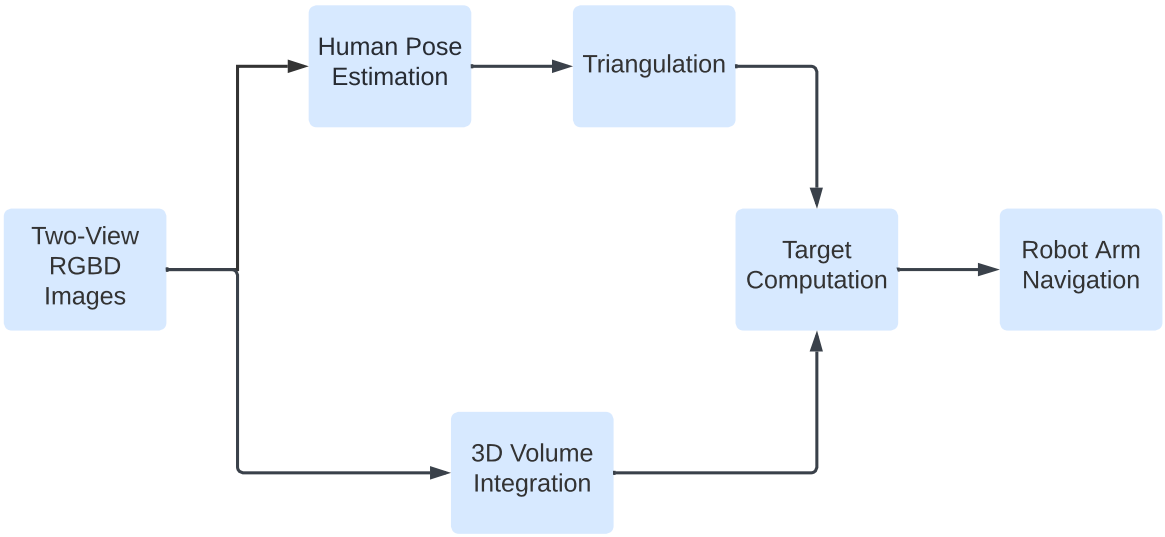


Figure 3. System pipeline.

Coordinate Transformation

There are four coordinate frames in our transducer positioning process. The robot base frame, the robot gripper frame, and the coordinate frames of the two cameras. The robot base coordinate frame is used as the global coordinate system. The transformation matrix from gripper to base $T^{base}_{gripper}$ can be obtained from robot system output with very high accuracy and precision, while the transformations from camera to robot base frame T^{base}_{cam} are obtained via hand-eye calibration.

Hand-eye Calibration

Hand-eye calibration technique is used to compute the transformation matrix from the camera coordinate frame to the robot base frame, denoted as T^{base}_{cam}. Since both cameras are stationary with respect to robot base, we adopt the eye-to-hand calibration setting. A calibration pattern is fixed on the robot gripper, and the robot is programmed to move across the field of view of the camera, while color pictures capturing the calibration pattern and the corresponding robot gripper poses are recorded.

Specifically, we use Apriltag [41] as the calibration pattern. This fiducial tag system allows us to directly obtain from the captured image the transformation from the tag coordinate frame to the camera coordinate frame. We denote this transformation as T^{cam}_{tag}, where

$$T^{cam}_{tag} = \begin{bmatrix} R^{cam}_{tag} & t^{cam}_{tag} \\ \mathbf{0}^T & 1 \end{bmatrix} \tag{1}$$

Overall, there are four transformation matrices involved in the calibration process, between robot base, camera, Apriltag and robot gripper. The relationship between these transformations is shown in Fig. 4, and can be expressed as

$$T_{tag}^{gripper} = T_{base}^{gripper}\ T_{cam}^{base}\ T_{tag}^{cam} \tag{2}$$

$T_{base2gripper}$ is computed as the inverse of $T_{gripper2base}$, which can be easily and accurately obtained via program output.

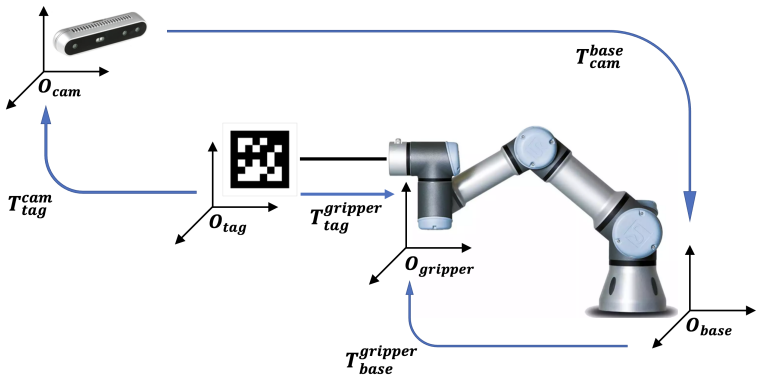


Figure 4. Coordinate frame transformation in eye-to-hand calibration

For eye-to-hand calibration setting, both transformations from tag frame to gripper frame and from camera frame to base frame do not change. For any two robot gripper poses, each with a distinct set of $T_{base}^{gripper}$ and T_{tag}^{cam}, the following relationship holds:

$$(T_{base}^{gripper})_1\ T_{cam}^{base}\ (T_{tag}^{cam})_1 = (T_{base}^{gripper})_2\ T_{cam}^{base}\ (T_{tag}^{cam})_2 \tag{3}$$

$$T_{cam}^{base}\ (T_{tag}^{cam})_1\ (T_{tag}^{cam})_2^{-1} = (T_{base}^{gripper})_1^{-1}\ (T_{base}^{gripper})_2\ T_{cam}^{base} \tag{4}$$

Let $A = (T_{base}^{gripper})_1^{-1}\ (T_{base}^{gripper})_2$, $B = (T_{tag}^{cam})_1\ (T_{tag}^{cam})_2^{-1}$, $X = T_{cam}^{base}$, where A and B are known. This becomes the $AX = XB$ problem where the closed-form least squares solution can be computed when A and B are measured in the presence of noise. We implemented the method proposed by Park et al. [43] to obtain estimates for the transformations from camera to robot base frame. In practice, we collected data from more than 20 different poses for each camera.

Pose Estimation

A deep learning based human pose estimation (HPE) model is deployed to extract keypoint features from the two-view color images of the subject. In our experiments, we would like to obtain accurate estimation for the locations of the two shoulders and the right hip of the subject in each image. To achieve that, we ensure that the images captured contain the upper torso of the subject from head to waist. To evaluate the effect of different pose estimation models on the overall system performance, we compared the performance using two Vision-Transformer (ViT) based model and one Convolutional-Neural-Network (CNN) based model.

ViTPose [57], a recent pose estimation framework based on vision transformer, has been included in our experiments. It follows the common top-down setting, where the pose estimator employs plain and non-hierarchical vision transformers as backbones to extract features from a given person instance, and a lightweight decoder for pose estimation. It also has great scalability, as the model size varies from 100M to 1B parameters, where the largest model represents a new state-of-the-art model. To speed up the runtime, a YOLOv3 model is used as person detector in the top-down HPE structure. Specifically, we selected the ViTPose-Base model that is pretrained on COCO dataset, and the ViTPose-large model with multi-task training on COCO+AIC+MPII+Crowdpose datasets, the later achieves state-of-the-art performance on all of the aforementioned datasets.

Also, we have tested our pipeline with the OpenPose model [6], which adopts the bottom-up approach and can achieve realtime multi-person 2D pose detection. In the first stage, a set of 2D confidence maps and a set of 2D vector fields of image size are produced. Each confidence map corresponds to a specific body part, indicating the probability of a pixel being the anatomical keypoint. The vector fields are known as Part Affinity Fields, where each one corresponds to a specific body limb, indicating the direction of the limb extension.

The performance of the HPE models on the MS COCO validation set is summarized in Table 1.

Table 1. Summary of HPE models tested in our experiments, including the design approach, train dataset and results on MS COCO validation set.

Model	Approach	Train Dataset	AP
OpenPose	Bottom-Up	COCO	61.8
ViTPose-B	Top-Down	COCO	75.8
ViTPose-L	Top-Down	COCO+AIC+MPII+CrowdPose	79.1

Triangulation

Triangulation is a classic Computer Vision technique used for 3D coordinates estimation from corresponding 2D coordinates in multi-view geometry [16]. Our goal is to compute the 3D coordinates of three keypoints (two shoulders and the right hip) in two-view images. We utilize the *triangulatePoints* function provided by the OpenCV library [5] to solve this problem, which is based on the Direct Linear Transformation (DLT) method [15].

The inputs of the triangulation algorithm include the projection matrices of the two cameras and the 2D pixel coordinates of the points in each view. The camera projection matrix $P \in \Re^{3\times4}$ is the product of camera intrinsic matrix $K \in \Re^{3\times3}$ and the extrinsic matrix $M = [R|t] \in \Re^{3\times4}$. Very accurate camera intrinsic parameters can be directly obtained using the *pyrealsense2* library, while the extrinsic parameters are provided by hand-eye calibration results. Note that the extrinsic matrix represents the transformation from robot base to camera, which is the inverse of the transformation from camera to robot base. By assuming that the pose estimation result is very accurate, we directly use the 2D keypoints of the body parts in two-view images as the triangulation input.

3D Volume Integration

3D volume integration problem is combining multi-view RGB-D images together given the camera poses. It is used for two purposes, surface normal estimation for robot pose computation and depth error compensation for triangulation result.

Ideally, when performing US scan, the probe mounted on the robot gripper should be perpendicular to the body surface, which is equivalent to being aligned with the surface normal at the point of contact.

We use the volume integration function provided by the *Open3D* library [62], which integrates each RGB-D frame incrementally into a volumetric representation using a truncated signed distance function (TSDF) [9, 37]. Besides the RGB-D data, the intrinsic and extrinsic information of the cameras is required as input. The output can be represented as point cloud format, from which the global coordinate and surface normal of each 3D scene point could be extracted.

The surface normal of the target point is estimated with a Nearest-Neighbor method. We assume that the point cloud is very dense so that the variation of surface normal between the target point and its nearest neighbor is infinitesimal. In practice, we discovered that for triangulation, the depth estimation error is usually significantly larger than width and height estimation errors. This is probably due to the relatively narrow baseline between the two cameras, whereas triangulation generally produces better result with wider baseline given accurately matched features [40]. As a result, due to the large deviation between depth (Z-axis) estimate using triangulation and 3D volume integration, we conduct the nearest neighbor search in terms of Euclidean distance on width and height dimensions (XY-plane) only. We refer to this metric as the Planar Euclidean Distance. Also, empirically we find out that the depth estimate from triangulation is usually deeper than the actual value. This could cause safety issue when the robot is driven too deep into the human body. To address this problem, we replace the depth estimate of target

point from triangulation with the depth value of its nearest neighbor in terms of Planar Euclidean Distance.

Denote the target point as T with 3D coordinate (X_T, Y_T, Z_T), the set of point cloud as A, and the nearest neighbor of T in A as P, the surface normal $\hat{n}_T$ and depth estimate Z_T of the target can be expressed as

$$\begin{aligned} &\hat{n}_T = \hat{n}_P,\ Z_T = Z_P \\ &\text{where } P = \arg\min_{a \in A} ||(X_a, Y_a) - (X_T, Y_T)||_2 \end{aligned} \tag{5}$$

Robot Target Pose Computation

The pose of robot gripper in robot base frame is expressed as a 6D vector (X, Y, Z, RX, RY, RZ), where the first three elements denote the position and the last three denote the orientation as a 3D rotation vector in angle-axis representation. The position of the target robot pose is the 3D target coordinate, and the orientation is the equivalent rotation vector of the rotation matrix that aligns the robot gripper to the target surface normal.

Target Position

We define an anatomical model with empirically determined ratio parameters to estimate the three target positions in a 3D setting, illustrated in Fig. 5. The model takes certain 3D keypoint locations as input, applies the ratio parameters by interpolating the input keypoints, and outputs the target 3D locations in the global frame.

Due to the scarce of training data, we have to use rule-based parametric models to compute target scan locations from 3D keypoint coordinates. We explicitly assume that all male subjects have identical proportions in terms of human anatomy. We use two parametric models for targets located on the front and side of human body, each model has two parameters that could either be predefined or regressed from collected data. The

front targets share the same model structure but not all the parameters.

Front Pose (Fig. 5a)

There are two scan targets located on the front of the human body, namely locations 1 and 2 in Fig. 1. For the human subject illustrated in Fig. 5a, we determine the target points T_1, T_2 from the shoulder locations X_1, X_2 and two ratio parameters r_1, r_2. Theoretically, T_1, T_2 should lie on the mid-clavicular line (MC), which is a line that origins from the halfway between sternoclavicular and acromioclavicular joints and parallel to the anatomic midline [36]. In practice, its origin is approximately at the quartile between the two shoulder keypoints. Therefore, we first locate X_3 on the line segment $\overline{X_1X_2}$ where the MC should pass through, with a ratio r_1 over the length of $\overline{X_1X_2}$. Then, we compute the equation of the line that passes through X_3, orthogonal to $\overline{X_1X_2}$, and parallel to the XY-plane of the global coordinate frame. The target location is computed with another ratio r_2 over the length of $\overline{X_1X_2}$. Specifically, T_1, T_2 share the same r_1 value but have different r_2 value. The previous computation assumes that the target has the same depth as X_3, which is not necessarily true. Again, by taking the advantage of the integrated point cloud from 3D volume integration, we update the depth value of the target using the Nearest Neighbor method with Planar Euclidean Distance.

Side Pose (Fig. 5b)

There is one scan target on the side of the human body, which is location 4 in Fig. 1, and it lies on the Anterior Axillary Line (AAx). It is very difficult to explicitly locate this line on each individual, yet based on empirical observations, we make the assumption that the AAx is parallel to the line connecting the right shoulder and right hip of the subject in 3D space. In this case, we denote X_2 as the position of right hip and T_4 as the target. The same interpolation method is used, with one distinct difference. We first locate X_3 on the line segment $\overline{X_1X_2}$ with a ratio r_1 over the length of $\overline{X_1X_2}$. Then we compute the equation of the line that passes through X_3, is orthogonal to $\overline{X_1X_2}$, and is

parallel to the XY-plane of the global coordinate frame. We assume the target lies on this line and compute its position with a ratio r_2, over the length of $\overline{\mathrm{X_1X_3}}$ instead of $\overline{\mathrm{X_1X_2}}$. This is because compared to the distance between X_3 to T_4, $\overline{\mathrm{X_1X_2}}$ is too large and the correlation between them might be weak. The depth value of the target is then updated in the same fashion.

The detailed steps of target computation are as follows

1. Compute direction vector $\vec{t_1}$ of the line X_1X_2

$$\vec{t_1} = \frac{X_2 - X_1}{||X_2 - X_1||_2} \tag{6}$$

2. Compute position of X_3 on line X_1X_2

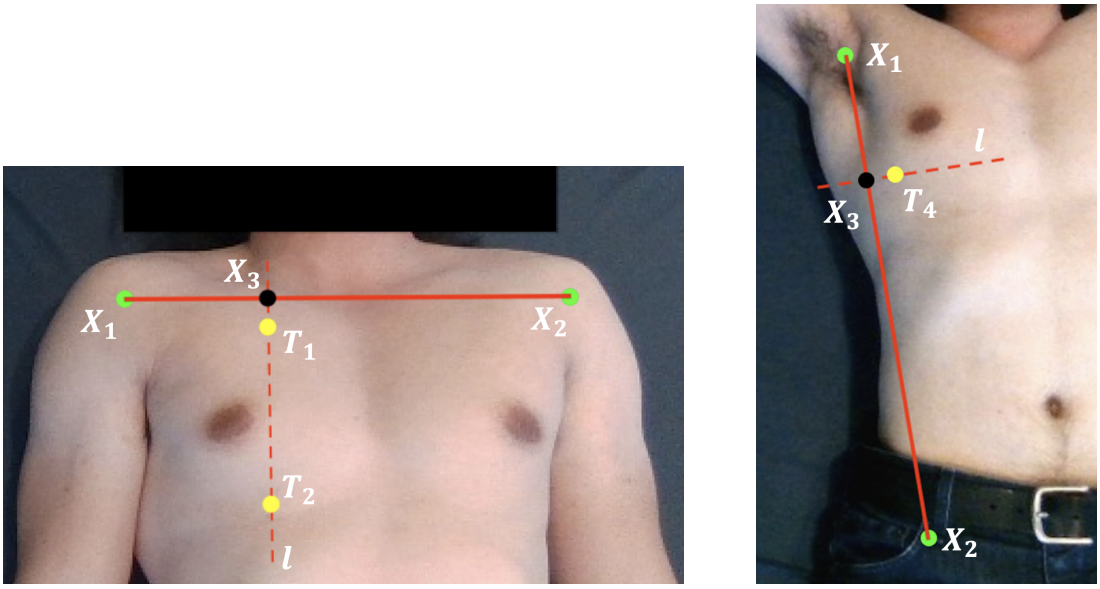
$$\vec{X_3} = X_1 + r_1(X_2 - X_1) \tag{7}$$

3. Compute direction vector $\vec{t_2}$ of the line perpendicular to X_1X_2 and parallel to the XY-plane of the global frame. Note that being parallel to a plane is equivalent to being orthogonal to its normal vector, and the normal vector of the XY-plane is just $\vec{n} = \begin{bmatrix} 0 & 0 & 1 \end{bmatrix}^T$

$$\vec{t_2} = \mathcal{N}(\begin{bmatrix} \vec{t_1}^T \\ \vec{n}^T \end{bmatrix}) \tag{8}$$

4. Compute the position of target on the line $l = X_3 + \lambda\vec{t_2}$

$$T = \begin{cases} X_3 + r_2\,\vec{t_2}\,||X_2 - X_1||_2 & \text{Front targets} \\ X_3 + r_2\,\vec{t_2}\,||X_3 - X_1||_2 & \text{Side target} \end{cases} \tag{9}$$



(a) Two front points. **(b)** One side point.

Figure 5. Target position estimation.

Parameter Optimization

With the collected data from our subjects, we are able to optimize the parameters of the models. The model for front targets is linear, and its parameters can be optimized through the least square method. However, the model for side target is non-linear, and we use Stochastic Gradient Descent (SGD) for optimization.

Since the ground truths are also obtained through triangulation algorithm from their pixel coordinates in each view, their depth values could be susceptible to the related systematic error. As a result, instead of minimizing the 3D Euclidean distance between the predicted targets and ground truths, we minimize the Planar Euclidean Distance between them. The optimization objective for both models could be expressed as:

$$r_1^*, r_2^* = \arg\min_{r_1 r_2 \in \Re} \frac{1}{N} \sum_{1}^{N} ||(X_{Pred}, Y_{Pred}) - (X_{GT}, Y_{GT})|| \tag{10}$$

Target Orientation

The target orientation can be expressed as the rotation from target frame to base frame $R_{tar}^{base} \in \Re^{3\times3}$, which can be expressed as $R_{tar}^{base} = R_{gripper}^{base} R_{tar}^{gripper}$. The matrix $R_{tar}^{gripper}$ is the rotation from the target frame to the gripper frame, aligning the target surface normal to the robot gripper, where the Z-axis of the robot gripper should point in the opposite direction of the target surface normal. The alignment algorithm was implemented based on [45], where trigonometry and normalization operations are avoided to improve stability and performance. The output rotation matrix $R_{tar}^{gripper}$ is then used to compute the overall orientation of the gripper, which is subsequently converted to a 3D rotation vector via the *pymath3d* [28] library.

Experiments

Data Collection

Our subject profile consists of 30 East Asian males with age between 20-27 year-old, height from 168 to 187 cm, and weight between 50 to 130 kg.

We first label the ground truth scan targets on the human subject. A Butterfly iQ portable US transducer probe is used to find the target scan points specified in [53] on individual subjects, where the probe is connected to an iPad which displays the real-time US image through the Butterfly iQ App. Compared to the Verasonic machine for obtaining high quality US images, the Butterfly probe is more convenient to operate while still accurate enough for the labeling task. The center point of the contact area between the transducer probe and the subject is marked as the ground truth scan target. The 3D coordinate of the ground truth is computed via triangulation, where the depth estimate is then updated using the Nearest Neighbor algorithm.

Each subject is asked to lie on the stretcher with two different poses. For US imaging on the front targets, the arms of the subject stay close to the torso, whereas for side targets the arms are raised with hands behind the head. Two sets of RGB-D images are captured for each subject from the two-view RealSense cameras. The pixel coordinate of the scan targets in the two-view images is then manually recorded.

Evaluation

Online evaluation is conducted by moving the robot to the target scan poses. Due to the limited range of motion of the Universal UR3e robot, many target poses could

not be reached especially for front targets. Also, it is very difficult to directly measure errors in width, height and depth during online evaluation with physical measurement tools. Therefore, we only report the results from offline evaluation.

For offline evaluation, the 3D coordinate of the ground truths is computed through triangulation. We then use the Nearest Neighbor algorithm to obtain the normal vector estimate and update the depth estimate of the ground truth coordinate. 3D Euclidean distance is used as the metric to evaluate the target position estimation error, while the target orientation error is expressed in terms of the error of normal vector estimation, which is the angle between the predicted vector and ground truth.

To the best of our knowledge, we are the first group to employ two-view RGB-D cameras for US scan target localization. To illustrate that our two-view setup yields more consistent 3D reconstruction result, we compare the outputs of two-view and single-view algorithms from the ground truth pixel coordinates. For single-view cameras, we use the provided API in RealSense SDK to deproject a pixel to its 3D coordinate, same as in [31]. We also update the depth estimate using the Nearest Neighbor algorithm for single-view results for fair comparison.

To accelerate the inference process of deep learning models, we run the person detection and human pose estimation models on GPU (NVIDIA GeForce RTX 3060, 6GB).

Parameter Optimization

We optimize the parameters for the target interpolation model using all of our collected data. We do not split the data into train and test sets because the sample size is still too small, and that our priority is to test the ability of the model to fit on the existing data.

The target interpolation model is linear for front targets while quadratic for side target, and both models are optimized based on the objective specified by eq. (10). We use a linear least square solver and SGD to compute the optimal parameters for front and

side targets respectively.

Analysis

We have observed that the OpenPose model occasionally produces faulty results, including the absence and wrong assignment of the right hip keypoint. Given the sample size of 30, OpenPose has two faulty results and the ViTPose models have none. This illustrates that ViTPose is superior than OpenPose in terms of robustness. Faulty results are subsequently excluded in the following evaluations.

We define the success of scan target localization as having a target position estimate within a certain error threshold. Fig. 6 shows the heatmap of success rates under increasing error thresholds for different HPE models with each scan target. Two sets of results are recorded for each setting, with default and optimized parameters for the scan target interpolation model. The parameter optimization leads to an increase on success rate in all settings, where the magnitude of the improvement depends on the deviation between the default and optimal parameters. Again, in the following discussion we only compare the optimized models for fairness.

Without faulty pose estimation, the improvement of HPE model performance has a visible positive impact on the target localization accuracy, especially with low error thresholds and challenging targets. ViTPose-L outperforms the other two models by a significant margin when the error threshold is no greater than 15mm. It can also be observed that target 4, locating on the side of the body, is the most challenging overall. In this case, the succecss rate of ViTPose-L is at least 10% higher than the others.

The absolute error distribution of position and orientation on each target is described in Fig. 7 and Fig. 8. All three models have similar performance on target 1 and 2 in terms of the range and median of error. However, on target 4 they tend to have larger error, especially for the OpenPose model. This further corroborates the impact of HPE model on localizing challenging targets.

Table 2. Ratio parameters for 3 targets and 3 HPE models before and after optimization. ViT-L = ViTPose-Large, ViT-B = ViTPose-Base, OP = OpenPose.

	Target 1				Target 2				Target 4			
HPE Model	Default	ViT-L	ViT-B	OP	Default	ViT-L	ViT-B	OP	Default	ViT-L	ViT-B	OP
r_1	0.300	0.292	0.294	0.296	0.300	0.276	0.282	0.291	0.350	0.344	0.354	0.342
r_2	0.100	0.123	0.124	0.120	0.550	0.577	0.584	0.579	0.100	0.108	0.089	0.052

Fig. 9 compares the 3D coordinate estimation results of ground truth scan targets using two-view and single-view RGB-D cameras. Across all samples, there is a systematic difference between the results from two single-view cameras, reaching a maximum magnitude close to 40mm. The magnitude of the difference in 3D coordinate estimate is correlated to the relative position of the cameras. In our setup, the two cameras share similar Y and Z positions in the global coordinate frame while separated along the X-axis for approximately 0.5m. In contrast, the position estimate using two-view triangulation is usually between the two single-view results. We therefore conclude that the result of 3D coordinate estimation is significantly affected by camera position for single-view RGB-D cameras, whereas this effect can be mitigated by adopting a two-view setting.

Parameter optimization is done on each model separately for each target, and a set of default parameters are used for all models before their individual optimization. The optimization result in Table 2 shows that the default parameter is not too far away from the optimum, and that the optimal values are generally similar for all models except for target 4, where r_2 of OpenPose is significantly smaller. This deviation occurs because hip keypoint estimation is less consistent for OpenPose in general, and that its shoulder estimation degrades significantly with the hands-behind-head body pose for side target scan.

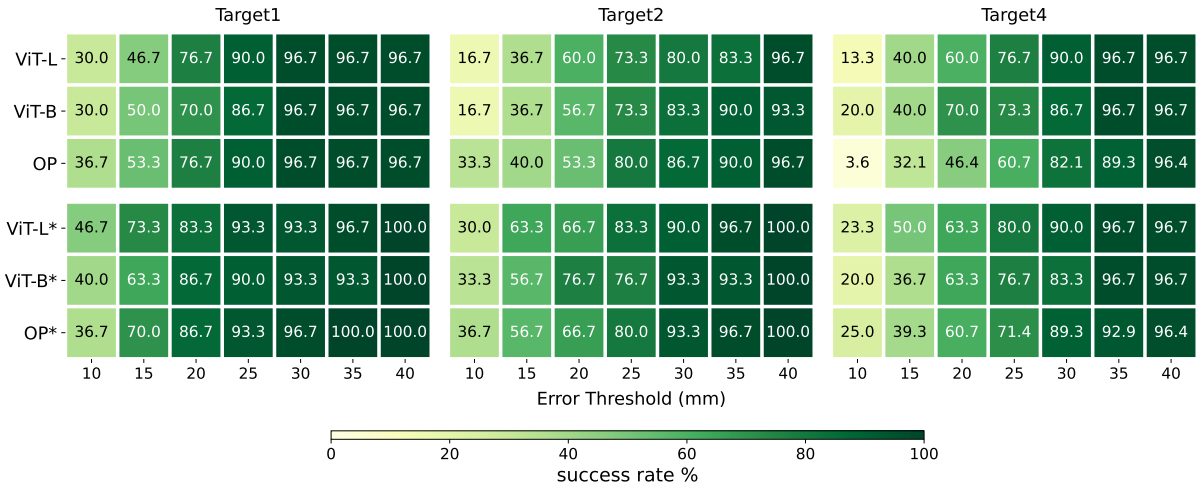


Figure 6. The localization success rate heatmap of 3 targets under increasing error threshold. ViT-L = ViTPose-Large, ViT-B = ViTPose-Base, OP = OpenPose, '$*$' means after parameter optimization.

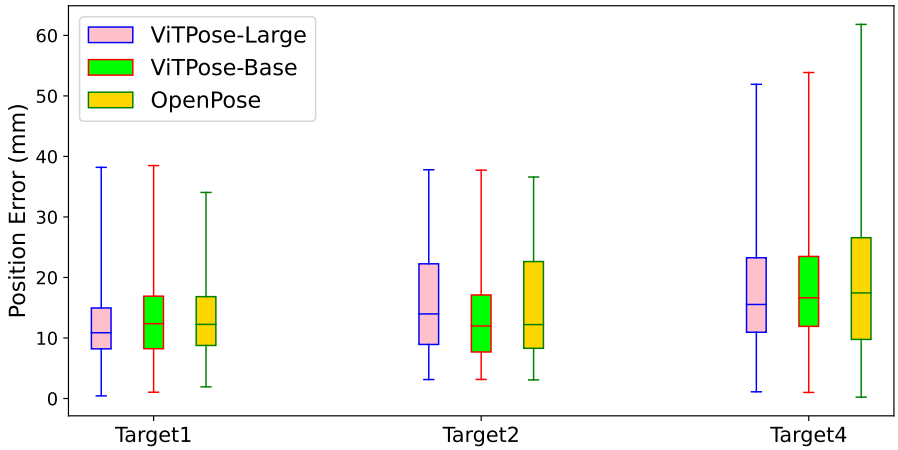


Figure 7. Position error distribution after parameter optimization.

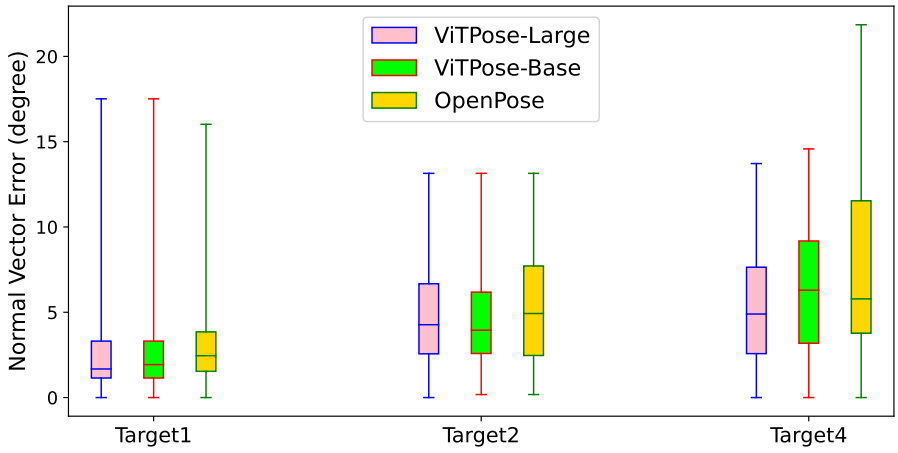


Figure 8. Normal vector error distribution after parameter optimization.

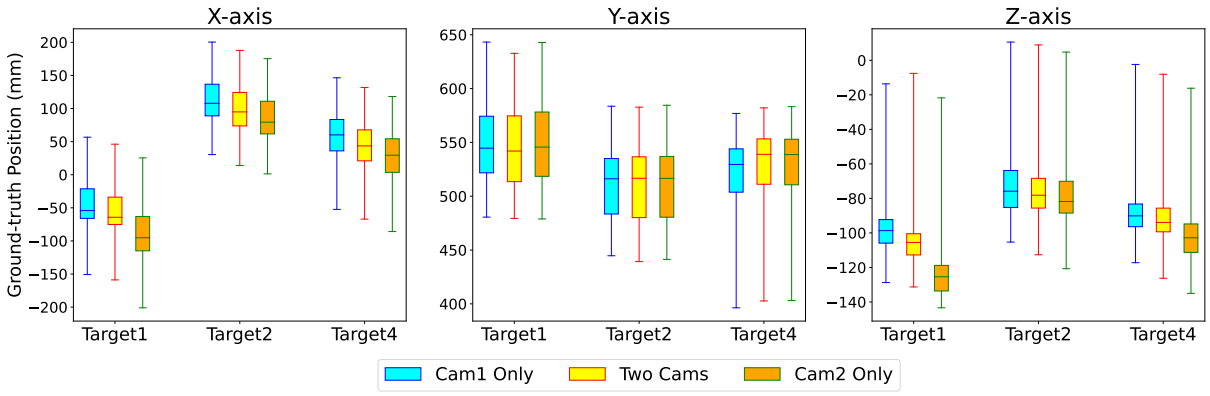


Figure 9. Ground-truth position distributions of using two-view and single-view RGB-D cameras.

Discussion

Source of Error

For our system which combines a variety of modules, error can arise and propagate between modules and eventually manifest in the system output. First, error from hand-eye calibration can affect the accuracy of triangulation and 3D volume integration. In parallel, the noise in depth image will affect the quality of 3D volume integration, normal vector estimation and the result of the nearest neighbor algorithm. Besides these, inaccurate pose estimation will also negatively affect triangulation result. Despite errors contained in the inputs from previous steps, triangulation also possesses an inherent error depending on the spatial relationship between the cameras and the scene. Finally, the success of the nearest neighbor algorithm heavily relies on the accuracy of triangulation and 3D volume integration, which undertakes the collective effect of previously mentioned systematic errors.

Limitation

The primary limitation of our solution lies in the scan target interpolation algorithm. By assuming all human subjects have identical proportions, we are able to propose a very simple model to effectively learn the relationship between the scan targets and other body part locations. However, human proportion can vary significantly among individuals, thus the model will struggle to closely fit on very diverse data.

Also, the sample size of 30 in our experiment is too small to train a model with

greater representation power. Empirically speaking, the size of data needs to be at least two magnitudes larger to achieve that.

Lastly, we did not integrate our system with other components of autonomous US imaging, therefore its effectiveness in complete autonomous lung US scan has yet to be verified. A practical autonomous US imaging system must be able to process sensor data and adjust probe movement accordingly in real-time. Currently, the runtime of our system is dominated by two processes, namely person detection (2.5 s) and 3D volume integration (2.3 s). Future improvement on runtime efficiency is necessary to address practical issues like body movement during scan.

Improvement

Improvements can be made by addressing the limitations from systematic error, model complexity, data collection and system integration. To reduce systematic error, sensors with higher quality could be deployed to reduce the noise in depth images. Also, we could adopt a more strict hand-eye calibration setting, such as using a customized, high-precision calibration pattern. Even though a very powerful HPE model has already been used, there is still possibility for improvements in the future.

Secondly, with large amount of data available, we will be able to relax the prior assumption on human anatomy and train a more complex model. One readily available method to achieve this is fine-tuning the existing HPE models like ViTPose, which is composed of an encoder for learning features and a decoder for the particular inference task. To fine-tune the model, the light weight decoder is re-trained on the data of the new task while the encoder parameters remain unchanged.

To acquire the data that could represent a more diverse subject profile, male subjects with a wider range of ethnicity and age should be included. Data from female subjects can also be collected in the future with strict ethical regulations.

Finally, to evaluate the performance of our system in a practical setting, we can incorporate a control module based on the feedback from US image and additional force sensors to adjust the US probe in real-time. Also, to improve the efficiency of our system, the runtime of person detection and 3D volume integration should be minimized. Separate processes could be employed to run these algorithms in parallel with the main program, and a fixed person detection bounding box could be imposed according to the position of the stretcher in each camera view.

Generalization

Among the wide range of US imaging types, lung US is a particularly challenging task. First, the variation on chest area is one of the largest among individuals, especially women. Second, the heavy presence of bones underneath the chest imposes additional obstruction to US imaging. Also, there are scan targets on the side of human body, which are proved to be harder to localize. In contrast, for other types of US imaging, such as thyroid, spine and fetus, there is usually less anatomical variation among individuals, along with universal visual cues that will help localization, and little obstruction from bone structure. As a result, we argue that our approach can be easily generalized to other types of US imaging tasks, given enough data from a diverse subject profile.

Conclusion

In this work, we propose a system that addresses the scan target localization problem for lung US imaging. This problem is critical for achieving fully autonomous US scan of all types and yet remains underexplored. As a purely vision based solution, our system adopts a multi-view stereo vision setting and incorporates various CV techniques, including human pose estimation, triangulation, and 3D volume integration. We also designed a scan target interpolation model based on explicit assumption on human anatomy. We test our method extensively on human subjects at a scale that is unparalleled compared to previous works. The average error of the position and normal vector estimate is approximately 15mm and 5° for all targets, which is accurate in terms of initial transducer positioning. For future work, dataset of larger scale and diversity needs to be collected to improve the complexity and generalization ability of the model. Workflow optimization can be done to reduce the runtime of the pipeline. Finally, integration with other components of autonomous US imaging system is necessary to test the performance in a practical setting. By qualifying for the challenging lung US scan, we believe our approach can be very well generalized to other types of US imaging tasks.

This thesis is currently being prepared for submission for publication of the material. Long, Jianzhi; Cai, Jicang; Al-Battal, Abdullah; Jin, Shiwei; Zhang, Jing; Tao, Dacheng; Nguyen, Truong. “A Vision-Based Approach to Scan Target Localization for Autonomous Lung Ultrasound Imaging”. The thesis author was the primary investigator and author of this paper.

Bibliography

[1] Mojtaba Akbari, Jay Carriere, Tyler Meyer, Ron Sloboda, Siraj Husain, Nawaid Usmani, and Mahdi Tavakoli. Robotic ultrasound scanning with real-time image-based force adjustment: quick response for enabling physical distancing during the covid-19 pandemic. *Frontiers in Robotics and AI*, 8:645424, 2021.

[2] Abdullah F Al-Battal, Yan Gong, Lu Xu, Timothy Morton, Chen Du, Yifeng Bu, Imanuel R Lerman, Radhika Madhavan, and Truong Q Nguyen. A cnn segmentation-based approach to object detection and tracking in ultrasound scans with application to the vagus nerve detection. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 3322–3327. IEEE, 2021.

[3] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition*, pages 3686–3693, 2014.

[4] Danilo Avola, Luigi Cinque, Alessio Fagioli, Gianluca Foresti, and Alessio Mecca. Ultrasound medical imaging techniques: a survey. *ACM Computing Surveys (CSUR)*, 54(3):1–38, 2021.

[5] Gary Bradski. The opencv library. *Dr. Dobb's Journal: Software Tools for the Professional Programmer*, 25(11):120–123, 2000.

[6] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299, 2017.

[7] Yucheng Chen, Yingli Tian, and Mingyi He. Monocular human pose estimation: A survey of deep learning-based methods. *Computer Vision and Image Understanding*, 192:102897, 2020.

[8] Jianning Chi, Ekta Walia, Paul Babyn, Jimmy Wang, Gary Groot, and Mark Eramian. Thyroid nodule classification in ultrasound images by fine-tuning deep convolutional neural network. *Journal of digital imaging*, 30(4):477–486, 2017.

[9] Brian Curless and Marc Levoy. A volumetric method for building complex models from range images. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, pages 303–312, 1996.

[10] Qi Dang, Jianqin Yin, Bin Wang, and Wenqing Zheng. Deep learning based 2d human pose estimation: A survey. *Tsinghua Science and Technology*, 24(6):663–676, 2019.

[11] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. Rmpe: Regional multi-person pose estimation. In *Proceedings of the IEEE international conference on computer vision*, pages 2334–2343, 2017.

[12] Raghavv Goel, Fnu Abhimanyu, Kirtan Patel, John Galeotti, and Howie Choset. Autonomous ultrasound scanning using bayesian optimization and hybrid force control. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 8396–8402. IEEE, 2022.

[13] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7297–7306, 2018.

[14] Mark A Haidekker. Medical imaging technology, 2013.

[15] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, ISBN: 0521540518, second edition, 2004.

[16] Richard I Hartley and Peter Sturm. Triangulation. *Computer vision and image understanding*, 68(2):146–157, 1997.

[17] Qinghua Huang, Yonghao Huang, Yaozhong Luo, Feiniu Yuan, and Xuelong Li. Segmentation of breast ultrasound image with semantic classification of superpixels. *Medical Image Analysis*, 61:101657, 2020.

[18] Qinghua Huang, Bowen Wu, Jiulong Lan, and Xuelong Li. Fully automatic three-dimensional ultrasound imaging based on conventional b-scan. *IEEE transactions on biomedical circuits and systems*, 12(2):426–436, 2018.

[19] Shaoli Huang, Mingming Gong, and Dacheng Tao. A coarse-fine network for keypoint localization. In *Proceedings of the IEEE international conference on computer vision*, pages 3028–3037, 2017.

[20] Elisee Ilunga-Mbuyamba, Juan Gabriel Avina-Cervantes, Dirk Lindner, Felix Arlt, Jean Fulbert Ituna-Yudonago, and Claire Chalopin. Patient-specific model-based segmentation of brain tumors in 3d intraoperative ultrasound images. *International journal of computer assisted radiology and surgery*, 13(3):331–342, 2018.

[21] Zhongliang Jiang, Matthias Grimm, Mingchuan Zhou, Javier Esteban, Walter Simson, Guillaume Zahnd, and Nassir Navab. Automatic normal positioning of robotic ultrasound probe based only on confidence map optimization and force measurement. *IEEE Robotics and Automation Letters*, 5(2):1342–1349, 2020.

[22] Muhammed Kocabas, Salih Karagoz, and Emre Akbas. Multiposenet: Fast multi-person pose estimation using pose residual network. In *Proceedings of the European conference on computer vision (ECCV)*, pages 417–433, 2018.

[23] Sven Kreiss, Lorenzo Bertoni, and Alexandre Alahi. Pifpaf: Composite fields for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11977–11986, 2019.

[24] Ching-Yen Lee, Tan-Loc Truong, and Pai-Chi Li. Automated conformal ultrasound scanning for breast screening. *Journal of Medical and Biological Engineering*, 38(1):116–128, 2018.

[25] Jiefeng Li, Can Wang, Hao Zhu, Yihuan Mao, Hao-Shu Fang, and Cewu Lu. Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10863–10872, 2019.

[26] Keyu Li, Yangxin Xu, and Max Q-H Meng. An overview of systems and techniques for autonomous robotic ultrasound acquisitions. *IEEE Transactions on Medical Robotics and Bionics*, 3(2):510–524, 2021.

[27] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[28] Morton Lind. pymath3d, 2018.

[29] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM transactions on graphics (TOG)*, 34(6):1–16, 2015.

[30] Jinlian Ma, Fa Wu, Tian'an Jiang, Qiyu Zhao, and Dexing Kong. Ultrasound image-based thyroid nodule automatic segmentation using convolutional neural networks. *International journal of computer assisted radiology and surgery*, 12(11):1895–1910, 2017.

[31] Xihan Ma, Ziming Zhang, and Haichong K Zhang. Autonomous scanning target localization for robotic lung ultrasound imaging. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 9467–9474. IEEE, 2021.

[32] Azeem Majeed, Mariam Molokhia, Bharat Pankhania, and Kaveh Asanati. Protecting the health of doctors during the covid-19 pandemic, 2020.

[33] Farhan Mohamed and C Vei Siang. A survey on 3d ultrasound reconstruction techniques. *Artificial Intelligence—Applications in Medicine and Biology*, pages 73–92, 2019.

[34] Ryu Nakadate, Jorge Solis, Atsuo Takanishi, Eiichi Minagawa, Motoaki Sugawara, and Kiyomi Niki. Out-of-plane visual servoing method for tracking the carotid artery with a robot-assisted ultrasound diagnostic system. In *2011 IEEE International Conference on Robotics and Automation*, pages 5267–5272. IEEE, 2011.

[35] Ryu Nakadate, Jorge Solis, Atsuo Takanishi, Motoaki Sugawara, Kiyomi Niki, and Eiichi Minagawa. Development of the ultrasound probe holding robot wta-1rii and an automated scanning algorithm based on ultrasound image feedback. In *ROMANSY 18 Robot Design, Dynamics and Control*, pages 359–366. Springer, 2010.

[36] C David Naylor, David G McCormack, and Stephen N Sullivan. The midclavicular line: a wandering landmark. *CMAJ: Canadian Medical Association Journal*, 136(1):48, 1987.

[37] Richard A. Newcombe, Shahram Izadi, Otmar Hilliges, David Molyneaux, David Kim, Andrew J. Davison, Pushmeet Kohli, Jamie Shotton, Steve Hodges, and Andrew W. Fitzgibbon. Kinectfusion: Real-time dense surface mapping and tracking. *2011 10th IEEE International Symposium on Mixed and Augmented Reality*, pages 127–136, 2011.

[38] Alejandro Newell, Zhiao Huang, and Jia Deng. Associative embedding: End-to-end learning for joint detection and grouping. *Advances in neural information processing systems*, 30, 2017.

[39] Guochen Ning, Xinran Zhang, and Hongen Liao. Autonomic robotic ultrasound imaging system based on reinforcement learning. *IEEE Transactions on Biomedical Engineering*, 68(9):2787–2797, 2021.

[40] Clark F Olson and Habib Abi-Rached. Wide-baseline stereo vision for terrain mapping. *Machine Vision and Applications*, 21(5):713–725, 2010.

[41] Edwin Olson. Apriltag: A robust and flexible visual fiducial system. In *2011 IEEE international conference on robotics and automation*, pages 3400–3407. IEEE, 2011.

[42] George Papandreou, Tyler Zhu, Nori Kanazawa, Alexander Toshev, Jonathan Tompson, Chris Bregler, and Kevin Murphy. Towards accurate multi-person pose estimation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4903–4911, 2017.

[43] Frank C Park and Bryan J Martin. Robot sensor calibration: solving AX= XB on the Euclidean group. *IEEE Transactions on Robotics and Automation*, 10(5):717–721, 1994.

[44] Alan M Priester, Shyam Natarajan, and Martin O Culjat. Robotic ultrasound systems in medicine. *IEEE transactions on ultrasonics, ferroelectrics, and frequency control*, 60(3):507–523, 2013.

[45] Inigo Quilez. Avoiding trigonometry in computer graphics, 2013.

[46] U Raghavendra, Hamido Fujita, Anjan Gudigar, Ranjan Shetty, Krishnananda Nayak, Umesh Pai, Jyothi Samanth, and U Rajendra Acharya. Automated technique for coronary artery disease characterization and classification using dd-dtdwt in ultrasound images. *Biomedical Signal Processing and Control*, 40:324–334, 2018.

[47] Kai Ritschel, Ioannis Pechlivanis, and Susanne Winter. Brain tumor classification on intraoperative contrast-enhanced ultrasound. *International Journal of Computer Assisted Radiology and Surgery*, 10(5):531–540, 2015.

[48] Septimiu E Salcudean, Gordon Bell, Simon Bachmann, Wen-Hong Zhu, Purang Abolmaesumi, and Peter D Lawrence. Robot-assisted diagnostic ultrasound–design and feasibility experiments. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 1062–1071. Springer, 1999.

[49] Ben Sapp and Ben Taskar. Modec: Multimodal decomposable models for human pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3674–3681, 2013.

[50] Ole Vegard Solberg, Frank Lindseth, Hans Torp, Richard E Blake, and Toril A Nagelhus Hernes. Freehand 3d ultrasound reconstruction algorithms—a review. *Ultrasound in medicine & biology*, 33(7):991–1009, 2007.

[51] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5693–5703, 2019.

[52] E Tegnander and SH Eik-Nes. The examiner's ultrasound experience has a significant impact on the detection rate of congenital heart defects at the second-trimester fetal examination. *Ultrasound in Obstetrics and Gynecology: The Official Journal of the International Society of Ultrasound in Obstetrics and Gynecology*, 28(1):8–14, 2006.

[53] David M Tierney, Joshua S Huelster, Josh D Overgaard, Michael B Plunkett, Lori L Boland, Catherine A St Hill, Vincent K Agboto, Claire S Smith, Bryce F Mikel, Brynn E Weise, Katelyn E Madigan, Ameet P Doshi, and Roman R Melamed. Comparative performance of pulmonary ultrasound, chest radiograph, and ct among patients with acute respiratory failure. *Critical Care Medicine*, 48(2):151–157, 2020.

[54] Heidi E Vanderpool, Elizabeth A Friis, Barbara S Smith, and Kenneth L Harms. Prevalence of carpal tunnel syndrome and other work-related musculoskeletal problems in cardiac sonographers. *Journal of occupational medicine*, pages 604–610, 1993.

[55] Adriana Vilchis, Jocelyne Troccaz, Philippe Cinquin, Kohji Masuda, and Franck Pellissier. A new robot architecture for tele-echography. *IEEE Transactions on Robotics and Automation*, 19(5):922–926, 2003.

[56] Jiahong Wu, He Zheng, Bo Zhao, Yixin Li, Baoming Yan, Rui Liang, Wenjia Wang, Shipei Zhou, Guosen Lin, Yanwei Fu, Yizhou Wang, and Yonggang Wang. Large-scale datasets for going deeper in image understanding. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1480–1485. IEEE, 2019.

[57] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose: Simple vision transformer baselines for human pose estimation. *arXiv preprint arXiv:2204.12484*, 2022.

[58] Zhoubing Xu, Qiangui Huang, JinHyeong Park, Mingqing Chen, Daguang Xu, Dong Yang, David Liu, and S Kevin Zhou. Supervised action classifier: Approaching landmark detection as image partitioning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 338–346. Springer, 2017.

[59] Cui Yang, Mingyao Jiang, Mianjie Chen, Maoqing Fu, Jianyi Li, and Qinghua Huang. Automatic 3-d imaging and measurement of human spines with a robotic ultrasound system. *IEEE Transactions on Instrumentation and Measurement*, 70:1–13, 2021.

[60] Yi Yang and Deva Ramanan. Articulated human detection with flexible mixtures of parts. *IEEE transactions on pattern analysis and machine intelligence*, 35(12):2878–2890, 2012.

[61] Song-Hai Zhang, Ruilong Li, Xin Dong, Paul Rosin, Zixi Cai, Xi Han, Dingcheng Yang, Haozhi Huang, and Shi-Min Hu. Pose2seg: Detection free human instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 889–898, 2019.

[62] Qian-Yi Zhou, Jaesik Park, and Vladlen Koltun. Open3D: A modern library for 3D data processing. *arXiv:1801.09847*, 2018.

[63] Andrzej Żytkowski, R Shane Tubbs, Joe Iwanaga, Edward Clarke, Michał Polguj, and Grzegorz Wysiadecki. Anatomical normality and variability: historical perspective and methodological considerations. *Translational Research in Anatomy*, 23:100105, 2021.