CVPR 2023 Ultrasonic Data Challenge Project Report

Nick Torenvliet
University of Waterloo
Waterloo, Ontario, Canada
ntorenvl@uwaterloo.ca

Yizhe Liu
University of Waterloo
Waterloo, Ontario, Canada
yihze.liu@uwaterloo.ca

Abstract

The CVPR Deep Learning in Ultrasound Image Analysis 3D Surface Mesh Estimation Challenge is to use some deep learning architecture to denoise volumetric industrial ultrasound images and provide estimates of the surface mesh of scanned object information.

1. Introduction

We are provided with 89 volumetric industrial ultrasound images, of variously connected pieces of steel pipe. Meshes for five of the images, manually cleaned by an experienced data analyst, were provided to serve as a pseudo ground truth. The objective is to use some architecture, presumably a deep learning architecture, and train a model to map the noisy ultrasound images to the underlying data, or surface, generating them. Model performance is measured using Chamfer distance and direct Hausdorff metric.

2. Rationale

We frame the problem as one involving denoising and edge finding. We base this on the description of the annotation process which involved manual cleaning of presumed noise, to make identifiable desired information in the dataset; where the desired information is the location of object surfaces. We investigated a number of approaches with state of the art architectures, for instance graph neural nets tailored for working with mesh data, but found them non-performant due to the ratio of model parameters vs. information in the available labelled data. This led us to consider smaller models with fewer parameters. Since the problem involves the detection of edges and surfaces we decided that Convolutional Neural Network(CNN) architectures, with their inductive bias suitable for computer vision related tasks and data modalities, were a natural choice. Since we are dealing with $3d$ datasets another seemingly natural choice was to use a $3d$ CNN. However we found that doing so effectively reduced the information available in our training data (due to the successive convolutions required) so that $3d$ CNN solution also exhibited inferior performance. Despite this fact the problem of edge finding in three dimension requires information in all three dimensions for optimal performance. Thus to respect a requirement to gain maximal information from the training data, while at the same time minimizing the over-all parameter count of the solution we opted to frame the problem as voxel classification in three dimensions given classification in three separate two dimensional planes.

3. Technical

3.1. Overview

We are given a dataset of ultrasonic images, consisting of training data X, labels Y, and unlabelled data R. To implement our solution we construct three subsets from the given training dataset X, of volumetric images and labels each consisting of slices of X in the xy, zx, and yz planes to give planar oriented datasets X_{xy}, X_{zx}, and X_{yz}. We model the solution as $v \sim P(v|\theta, v_{xy}, v_{zx}, v_{yz})$. Where the draw of random variable v is implemented as a voting $3d$ CNN V with parameters θ. The random variable conditionals are modelled as $v_{xy} \sim P(v_{xy}|\phi_{xy}, K, X_{xy})$, $v_{zx} \sim P(v_{zx}|\phi_{zx}, K, X_{zx})$, $v_{yz} \sim P(v_{yz}|\phi_{yz}, K, X_{yz})$, where the draws of random variables v_{xy}, v_{zx}, and v_{yz} are implemented as three separate $2d$ CNN V_{xy}, V_{zx}, and V_{yz} with parameters ϕ_{xy}, ϕ_{zx}, and ϕ_{yz} respectively. $K > 0$ is odd and determines the number of side by side planar slices used to make inference on a central slice. We implement K as the number of channels in V_{xy}, V_{zx}, and V_{yz}.

3.2. Data Preparation

To use the labelled training data we convert it to $3D$ occupancy fields that maintain the locations from $(0, 0, 0)$ to $(768 * 0.49479, 768 * 0.49479, 1280 * 0.3125)$ of the original meshes and have the same resolution $(768, 768, 1280)$ as the ultrasonic image data.

The conversion is based on center of voxel distance to the nearest mesh surface. If a voxel center

is within half of the diagonal distance across a voxel ($\sqrt{0.49479^2 + 0.49479^2 + 0.3125^2}/2$) of a mesh surface, its occupancy is set to 1 and otherwise 0. Fig. 1 show the results of converting the mesh for scan_001 with z axis pointing inward.

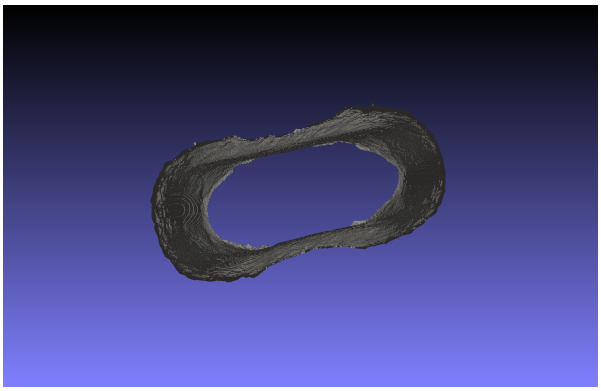

Figure 1. Occupancy Grid of Scan 001

3.3. 2D CNN Structure

CNN V_{xy}, V_{zx}, and V_{yz} are K channel $2d$ UNet CNN [1]. Each of the $2d$ UNet CNN trains on its own planar oriented datasets so that V_{xy}, V_{zx}, and V_{yz} train on X_{xy}, X_{zx}, and X_{yz} respectively; as per the models given above. We set $K = 5$ so that inputs to each $2d$ UNet CNN consist of five adjacent stacked slices from their respective dataset. Each of the $2d$ UNet CNN uses 5 down-sampling and 5 up-sampling blocks to predict the occupancy of the middle (or third) slice of each input. The $2d$ UNet CNN use the Binary Cross Entropy loss function. We train V_{xy}, V_{zx}, and V_{yz} on the labeled data, holding back some portion for validation, for $50 - 100$ epochs as directed by early stopping criteria. After training, a forward pass of the data from one of the ultrasound images is run, and predictions are stacked to obtain three $(768, 768, 1280)$ predicted occupancy fields as $P_{xy} = V_{xy}(X_{xy})$, $P_{zx} = V_{zx}(X_{zx})$, and $P_{yz} = V_{yz}(X_{yz})$. Ground truth and predicted slices in each plane are shown in Figure 2.

3.4. Voter: 3D CNN Structure

Voting $3d$ CNN V aggregates information from all 3 slicing directions as given by P_{xy}, P_{zx}, and P_{yz}. V is implemented as a $3d$ UNet [2] which takes input of dimension $(3, N, N, N)$ and predicts an (N, N, N) occupancy field. We choose $N = 64$ due to resource constraints. To form a dataset for training V we stack P_{xy}, P_{zx}, and P_{yz} to obtain data T of dimension $(3, 768, 768, 1280)$. Dataset X_v is constructed as the set of $12 * 12 * 20$ non-overlapping cuboids of dimension $(3, 64, 64, 64)$ in T, and we use the labels provided for ground truth. We train V on X_v, holding back

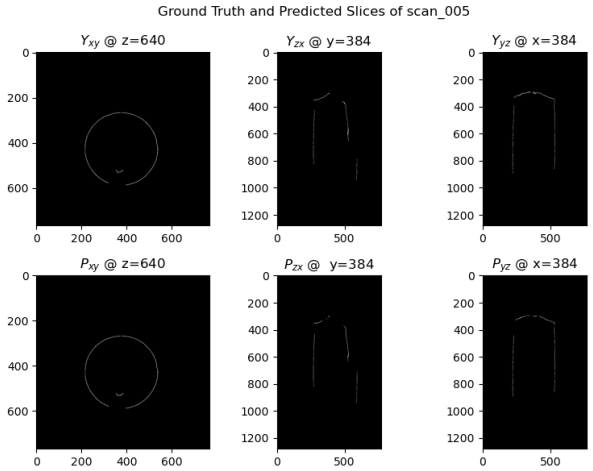

Figure 2. Prediction and Ground Truth of scan_005

some portion for validation, for $50 - 100$ epochs as directed by early stopping criteria. After training, a forward pass of the P_{xy}, P_{zx}, and P_{yz} associated with one of the ultrasound images is run, and predictions are stacked to obtain a predicted occupancy field P with dimensions $(768, 768, 1280)$. Figure 3 shows ground truth, V_{xy} prediction, and V prediction for two X_{xy} slices. We observe during our experimentation that V significantly decreases the occurrence of voxel miss-classification; on ultrasound image scan_005 V miss-classified 126517 while 133609 were produce by V_{xy}.

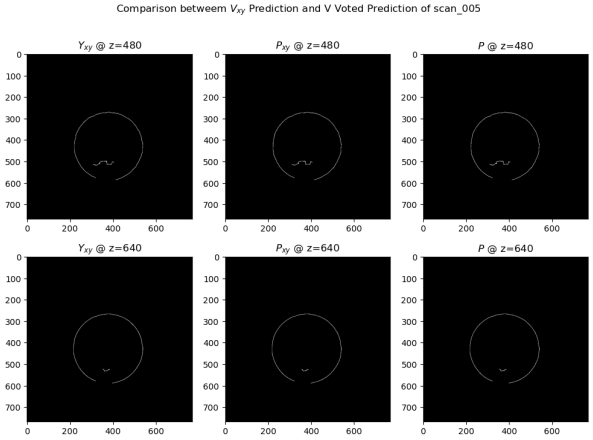

Figure 3. Comparison of scan_005

3.5. Evaluation

We convert the output occupancy grids back to point cloud by taking the center points of each occupied voxel. Then the point cloud is evaluated using the provided evaluation script.

We first show the quantitative result on scan_001 where

we have the ground truth in Table 1.

Metrics	Score
F Score (0.1)	$2.56 * 10^{-5}$
F Score (0.45)	0.0058
F Score (1.0)	0.0260
Chamfer Distance	1506.97
Direct Hausdorff	53.79
Mean Surface Distance	816.40
Residual Mean Square Distance	1254002

Table 1. Scan_001 Results

Figure 4 show the qualitative results on inference on scan_007 demonstrating model performance on out of distribution data, and suggesting reasonable generalization.

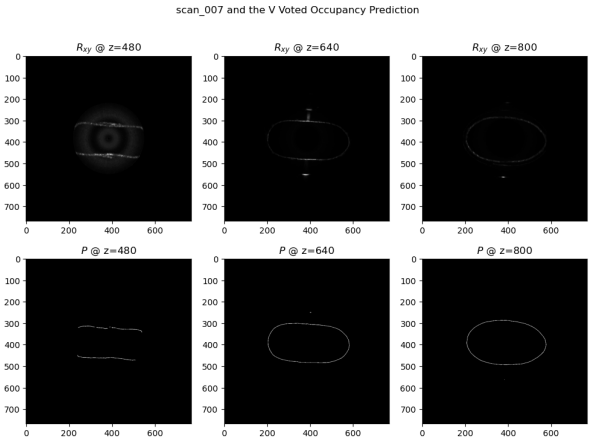

Figure 4. Scan_007 and the Occupancy Prediction

References

[1] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015. 2

[2] Özgün Çiçek, Ahmed Abdulkadir, Soeren S. Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: Learning dense volumetric segmentation from sparse annotation, 2016. 2