Final Report of 2023 DAIS Data Analytics Competition

Example 2: Surface Anomaly Detection from 3D point cloud data

Team member 1: **Xinchao Liu** (main contact)

Email: xl037@uark.edu

Affiliation: Department of Industrial Engineering, University of Arkansas

Team member 2: **Guanzhou Wei**

Email: gwei@uark.edu

Affiliation: Department of Industrial Engineering, University of Arkansas

March 6, 2023

1 Introduction

Our problem is *Example 2: Surface Anomaly Detection from 3D point cloud data*. The framework of this work is shown in Figure 1, and all steps in Figure 1 are implemented in Python. The first step is to load the unstructured point cloud data, then we segment the point cloud into two sub-regions including the reference surface region and the anomaly surface region. The ***tensor voting*** based segmentation algorithm is used to extract useful local geometry information by aggregation of spatial information from the neighborhood [Du et al., 2022]. Then a proposed ***untrained*** method, which can deal with a single sample without requiring additional data [Tao et al., 2022], is used for the segmentation of 3D point cloud samples. Feature engineering (or called feature extraction) is conducted on different sub-regions by exploiting descriptive statistics and advanced distance metrics including Wasserstein Distance, Chamfer Distance, and Hausdorff Distance. The pre-processing of data, e.g., the normalization of features and the balancing (we use SMOTE) of class distribution, are also performed before training the binary classifier (e.g., ***interpretable*** XGboost). Finally, the ***interpretable*** binary classifier is applied for surface anomaly classification prediction.

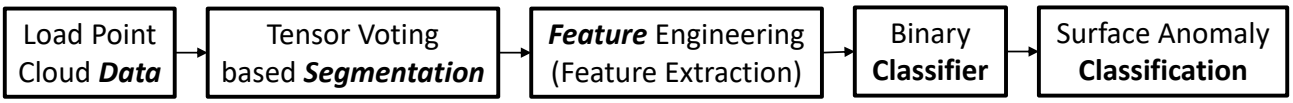

Figure 1: The framework of this work

2 Tensor-Voting based Segmentation for Reference Surface and Anomaly Surface

Before using tensor voting method to calculate all the local features, we can use point cloud *denoising* method to denoise the whole point cloud data. Although the *denoising* for the given dataset has been prepared, we find it is unnecessary because these data are not noisy after pre-checking.

2.1 Tensor Voting (for local point statistics)

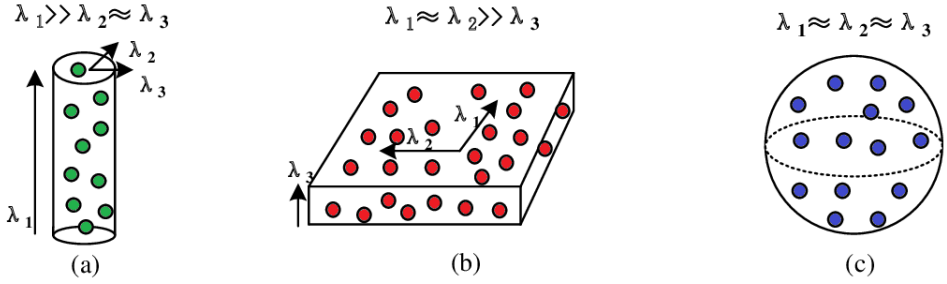

Figure 2: The saliency features including (a) L: Linearness, (b) P: Planarity, and (c) S: Sphericity

After performing the tensor voting approach for all 3D points, the saliency features shown in Figure 2 can be captured by the decomposition into principle components of the covariance matrix of the 3D points position. Let N denote the number of neighbors (points) from the support region of each point $X_i = (x_i, y_i, z_i) \in \mathbb{R}^3$. The symmetric positive definite covariance matrix for the set of N neighbors with $\overline{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$ is $\frac{1}{N}\sum_{i=1}^{N}(X_i - \overline{X})(X_i - \overline{X})^T$. Then this matrix is decomposed into principal components ordered by decreasing eigenvalues, i.e., $\lambda_1 \geq \lambda_2 \geq \lambda_3$. Finally, the salinency features can be captured by manipulating the eigenvalues $\lambda_1, \lambda_2, \lambda_3$, as shown by the top lines of subfigures in Figure 2.

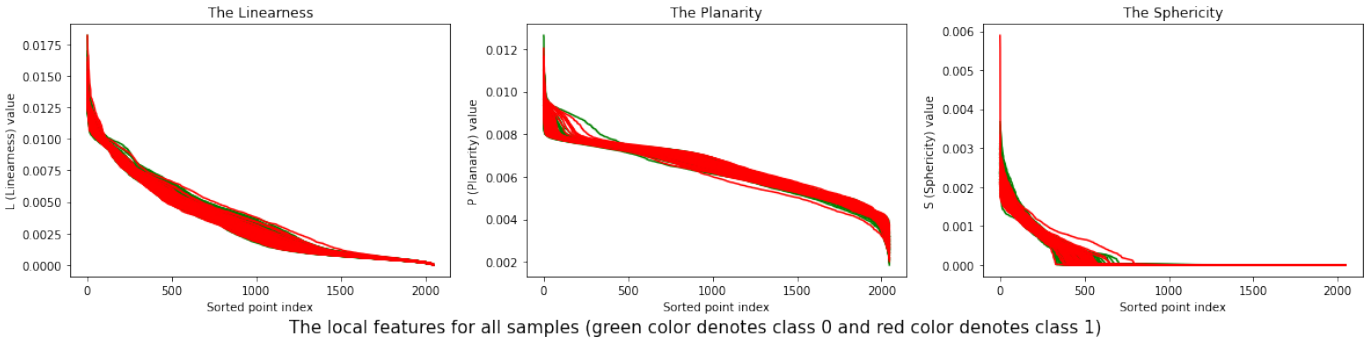

Figure 3: The extracted local features for all (i.e., 190) point cloud samples

Justification: The node-wise physical features (including L, S, and P), which indicate physical information, can help us get physical-interpreted segmentations, such as the reference surface and the anomaly surface. For example, the sub-figure on the right, i.e., the Sphericity, in Figure 3, clearly indicates the division of plane and non-plane surface around the 500th sorted point. If we directly apply clustering methods (such as k-means and DBSCAN (Density-based spatial clustering of applications with noise)) to the coordinates, the segmentations will lack the physical interpretations, e.g., the reference planes and anomalies. Moreover, the direct clustering of the coordinates is not stable or robust, when the shape or the junction saliency is changed. Based on the node-wise physical features (including L, S, and P), we can incorporate different kinds of clustering methods to achieve physical-interpreted features for classification purposes.

2.2 Gaussian Mixture Model (GMM)

Based on the tensor voting results of node-wise saliency features, references have proposed some *untrained* methods to decide the inferred point-type set, such as the piecewise-linear spline with free knots method [Du et al., 2022] and the novel Bayesian network with designed structure [Tao et al., 2022]. In this work, we proposed a *untrained* method using Gaussian Mixture Model (GMM) based on tensor voting results. One unsupervised segmentation sample is given in Figure 4.

GMM has the advantage of providing estimates of the probability that each data point belongs to each sub-regions of the point cloud. The saliency feature distribution is learned by GMM using EM algorithm. For each sub-region, the resulting density probability model is the sum of n_g Gaussians with weight, mean, and covariance matrices $\{(\omega_i, \mu_i, \Sigma_i)\}_{i=1,\ldots,n_g}$. The proposed *untrained* method can identify most of the reference surface points while only segmenting the point cloud into two sub-regions, i.e., the reference surface and the anomaly surface. During the segmentation, SVD is employed to fit the reference surface to further fine the segmented sub-regions.

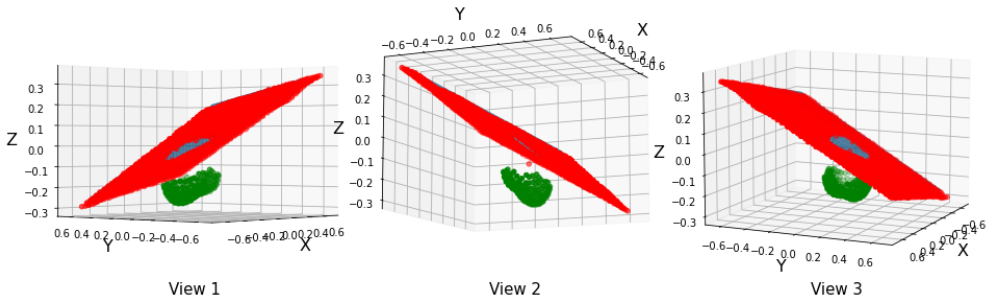

Figure 4: The illustration of our unsupervised segmentation results (sample #73)

3 Feature Engineering

We extracted features using descriptive statistics and advanced distance metrics. See Table 1,

Table 1: Extracted features including descriptive statistics and advanced distance metrics

Feature	Definition	Physical interpretations
$f_{0,k}$	Mean of the vertical distance to the reference surface	Characterization of distance distribution of surface anomalies in 3D
$f_{1,k} \sim f_{6,k}$	Quantiles of the vertical distance to the reference surface	Characterization of distance distribution of surface anomalies in 3D
$f_{7,k} \sim f_{9,k}$	Mean of the coordinates of anomaly surface nodes	Characterization of anomaly distribution in 3D
$f_{10,k} \sim f_{12,k}$	Standard deviation of the coordinates of anomaly surface nodes	Characterization of anomaly distribution in 3D
$f_{13,k} \sim f_{14,k}$	Size (length and width) of the dent	Characterization of anomaly shapes
$f_{15,k}$	Percentage of the anomaly surface nodes	Characterization of anomaly size
$f_{16,k} \sim f_{18,k}$	Mean of the coordinates of reference surface nodes	Characterization of distance distributions of artifact surface in 3D
$f_{19,k} \sim f_{21,k}$	Standard deviation of the coordinates of reference surface nodes	Characterization of distance distributions of artifact surface in 3D
$f_{22,k}$	Mean distance of nodes around surface (to the fitted surface)	Characterization of outliers
$f_{23,k} \sim f_{26,k}$	Mean and standard deviation of **Hausdorff distance** to Class 0 and 1	Characterization of discrepancy between 3D point sets
$f_{27,k} \sim f_{30,k}$	Mean and standard deviation of **Chamfer distance** to Class 0 and 1	Characterization of discrepancy between 3D point sets
$f_{31,k} \sim f_{34,k}$	Mean and standard deviation of **Wasserstein distance** to Class 0 and 1	Characterization of discrepancy between the distributions of 3D point sets

3.1 Descriptive Statistics

We used descriptive statistics features (including the shape and percentage of anomalies, the distribution of different sub-regions (i.e., k), the outliers, and the distribution of pairwise distances).

Descriptive statistics have the advantage of dealing with unequal point sizes and we do not have registration issues. All descriptive statistics features are corresponding to $f_{0,k} \sim f_{22,k}$ in Table 1.

3.2 Wasserstein Distance, Chamfer Distance, and Hausdorff Distance

Apart from the descriptive features that can be extracted from the related statistics, there may exist some *fine-grained* differences between the two anomaly point clouds. Here we also include three advanced metrics (Wasserstein, Chamfer, and Hausdorff distances) to characterize the discrepancy between the two anomaly point clouds. Let P and Q as two point clouds, which are two sets containing 3D points in $\mathbb{R}^3$. The two elements p and q represent two points in P and Q respectively (i.e., $p \in P \subset \mathbb{R}^3$, $q \in Q \subset \mathbb{R}^3$). The following three equations give how to calculate the Wasserstein, Chamfer, and Hausdorff distance between two point clouds P and Q, respectively:

$$
\begin{aligned}
d_{\mathrm{W}}(\mathbb{P}_P, \mathbb{P}_Q) &= \inf_{\mathbb{Q} \sim (\mathbb{P}_P, \mathbb{P}_Q)} \mathbb{E}_{\mathbb{Q}} \left(||X_P - X_Q||_{p'} \right) \\
d_{\mathrm{H}}(P,Q) &= \max \left\{ \sup_{p \in P} \left(\inf_{q \in Q} ||p-q||_2^2 \right), \sup_{q \in Q} \left(\inf_{p \in P} ||p-q||_2^2 \right) \right\} \\
d_{\mathrm{CD}}(P,Q) &= \max \left\{ \frac{1}{|P|} \sum_{p \in P} \min_{q \in Q} ||p-q||_2^2, \frac{1}{|Q|} \sum_{q \in Q} \min_{p \in P} ||p-q||_2^2 \right\},
\end{aligned}
\tag{1}
$$

where $d_{\mathrm{W}}, d_H, d_{\mathrm{CD}}$ denote the Wasserstein, Hausdorff, Chamfer distances respectively, $X_\cdot$ is a random vector following the distribution $\mathbb{P}_\cdot$, $\mathbb{Q}$ denotes the joint distribution of (X_P, X_Q), $||\cdot||_{p'}$ is the p'-norm, and $|\cdot|$ denotes the set size. The features are enumerated by $f_{23,k} \sim f_{34,k}$ in Table 1.

4 Balancing, Training, and Anomaly Classification Prediction

The class imbalance problem is handled by SMOTE method. Binary classifiers are compared in Table 2. For an example of hyperparameter tuning, when tuning the XGboost model, we performed GridSearchCV on max_{depth}, $min_{childweight}$, $eval_{error}$, reg_{alpha}, etc.

According to Table 2, we choose XGboost as our final binary classifier because it not only has good overall performance (considering all evaluation metrics) but also has great interpretability shown in Figure 5. Note that the performance will be improved with more data fed into the model.

Table 2: Comparisons between classifiers for 100 random splits on the testing dataset (20%)

Metrics	XGboost[a]	RF	LogitR	SVM	LightGBM	PointNet[b]
Avg. ACC	0.7329	0.7424	0.6811	0.6379	0.7263	0.6634
Avg. Precision	0.6155	0.6782	0.5268	0.4836	0.6178	-
Avg. Recall	0.5006	0.4397	0.5826	0.7123	0.5241	-
Avg. F1-score	0.5308	0.5119	0.5399	0.5503	0.5374	-

[a] XGboost is finally selected due to its relatively large average ACC and average F1-score.

[b] The PointNet model cannot be trained well with small dataset, while it always predicts class 0.

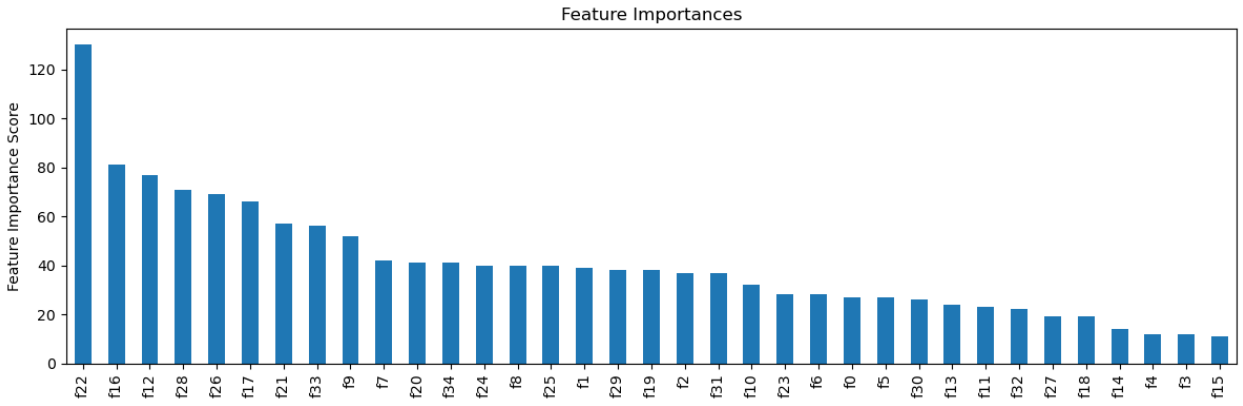

Figure 5: **Interpretability**-The calculated feature importances using XGboost by cross-validation.

5 Novelty and Contributions

- We proposed a novel *untrained* method that combines tensor voting and GMM model for unsupervised segmentation of unstructured 3D point cloud.
- We proposed descriptive statistics features for unequal point size within 3D anomalies.
- Advanced distance features (e.g., Wasserstein) are developed for depicting 3D anomalies.
- We solved the class imbalance problem using SMOTE for 3D point cloud classification task.
- We proposed a method to align the number of points (e.g., for Wasserstein and PointNet).

6 Summary

- The tensor voting method is very important for anomaly detection and classification of 3D point cloud data. We proposed a novel framework coupling tensor voting with GMM model.
- To the best of our knowledge, we are the first to simultaneously consider descriptive statistics features and advanced distance features for anomaly classification of 3D point cloud data. Our framework considering both equal and unequal point size can embrace many different classifiers, even the PointNet deep learning classifier if we have enough training data.

References

Juan Du, Hao Yan, Tzyy-Shuh Chang, and Jianjun Shi. A tensor voting-based surface anomaly classification approach by using 3d point cloud data. *Journal of Manufacturing Science and Engineering*, 144(5), 2022.

Chengyu Tao, Juan Du, and Tzyy-Shuh Chang. Anomaly detection for fabricated artifact by using unstructured 3d point cloud data. *IISE Transactions*, (just-accepted):1–29, 2022.