Identifying and Analyzing Data Quality Issues in Autonomous Driving Datasets

INFO 5731 - Computational Methods for Information Systems – Section 020

Aiswarya Goriparthi
11676073

Nitya Vattam
11690948

Vinuthna Reddy Vintha
11755782

Shanmukha varma kothapalli
11792494

Bhavya Madhuri Devarakonda
11698719

Pavani Kommineni
11648943

Contents

1 Introduction

1.1 Background

Autonomous vehicles (AVs) development represents a great leap forward in modern advanced transportation technology. However, the security and dependability of these systems are highly dependent on the data quality used for training and validation. Datasets such as KITTI, Waymo, and nuScenes give important multi-modal sensor data for the development of AV models. These datasets frequently have significant data quality issues, like labeling inaccuracies, insufficient data, misaligned sensors, and duplicate annotations.

These issues with quality challenges pose risks, as they can lead to incorrect object detection, bad decision-making, and , and ultimately drive dangerously in AV systems. This project focuses on analyzing and classifying these data quality (DQ) issues within AV datasets, putting emphasis on correctness, consistency, completeness, and redundancy. The aim is to reduce these issues to increase the reliability and security of autonomous driving systems.

1.2 Research Purpose

The purpose of this research is the systematic identification, analysis, and classification of data quality issues present in autonomous driving datasets, especially focusing on the KITTI and Waymo datasets. The research goal is to develop an all-encompassing framework that takes into account aspects like data quality, that includes correctness, consistency, completeness, and redundancy. By doing so, the project aims to develop the overall reliability and safety of AV systems. Additionally this research intends to give practical solutions for identifying and resolving these data quality concerns.

1.3 Significance

Improving the safety and effectiveness of autonomous systems by addressing data quality in AV datasets is important. Inaccuracies, misalignments, and missing data can lead to dangerous consequences like faulty predictions and unsafe behaviors, leading to severe consequences on the road. This research sets itself apart from prior studies by addressing multiple aspects of data quality, instead of just focusing on isolated issues like labeling errors or sensor calibration. By creating methods to detect and fix these issues, the research provides both academic advancements and practical implementations.

1.4 Research Questions:

(1) What kinds of data quality issues are present in AV datasets, especially in the KITTI dataset?
(2) How do data quality issues are different from across datasets, like KITTI and Waymo?
(3) How can multi-modal sensor data (e.g., lidar, radar) be used to identify mistakes in labeling and annotations?
(4) What practical frameworks can be used to increase the overall quality of AV datasets?

1.5 Research Methods

The research will use both manual and automated methods to find and examine data quality issues. Manual validation will have visual inspection of sensor data and annotations whereas automated techniques will have the development of scripts to evaluate error rates, find inconsistencies, and get missing data. Data cleaning and standardization methods, like interpolation for missing data and values and deduplication of redundant parts of data, will be applied to increase dataset quality. The results will be contrasted across multiple datasets like KITTI and Waymo to find common trends and dataset-specific issues.

2 Related Work

Recent research in the field of autonomous driving has led to the identification of several challenges related to the AV datasets quality. Many key studies have explored several aspects of data quality, which includes sensor calibration, labeling inaccuracies, and redundancy management.

(1) **Smith et al. (2021)** [1]: Analysed misalignment of sensors in AV datasets, suggesting methods for correcting calibration problems between camera and lidar data.
(2) **Liu and Zhao (2020)** [2]: Examined deep learning techniques to identify and rectify labeling errors in the KITTI dataset, focusing on object detection models.
(3) **Brown and Kim (2022)** [3]: Investigated the consistency of multi-modal datasets, especially the alignment between camera and lidar sensors across time.
(4) **Perez et al. (2019)** [4]: Addressed the issue of redundancy within AV datasets, proposing a framework for finding and eliminating unnecessary data points to increase training efficiency.
(5) **Wang et al. (2021)** [5]: Proposed a methodology for assessing the completeness of AV datasets, concentrating on missing data and incomplete annotations, while Garcia et al.
(6) **Garcia et al. (2020)** [6]: Developed statistical models to detect errors in large-scale datasets used for object detection tasks.
(7) **Martinez and Chen (2023)** [7]:
Studied the effects of wrong annotations on object detection models, coming to the conclusion that even small labeling errors could lead to major degradation in performance.
(8) **Park et al. (2019)** [8]:
Noted the challenge to maintain temporal consistency across time-series data, important for tasks such as trajectory prediction.
(9) **Henderson et al. (2022)** [9]:
Introduced techniques for data duplication, aiming to decrease the redundancy present in large AV datasets,
(10) **Nguyen et al. (2023)** [10]: Proposed methods for finding errors in 3D bounding boxes.

These studies give thorough strategy for the comprehensive approach taken in this research, which addresses all aspects of data quality—correctness, consistency, completeness, and redundancy—across various datasets.

3 Methodology

The methodology is structured to systematically answer the research questions, using both manual inspections and automated processes to identify data quality problems.

3.1 Dataset Selection and Exploration

The primary dataset used in this study is the KITTI Object Detection Dataset, a popular benchmark in autonomous driving research. KITTI gives pictures, lidar scans, and ground-truth annotations (e.g., 3D bounding boxes). These annotations are examined for accuracy, completeness, and consistency across frames and sensor modalities.

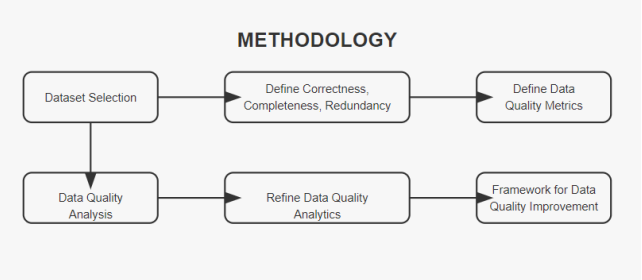

Figure 1: Overview of the methodology for detecting and categorizing issues of data quality in AV datasets.

The study incorporates the Waymo Open Dataset known for its large volume and complexity, for comparison analysis.

3.2 Defining Data Quality Metrics

The analysis of data quality is served by the following measures:

- **Correctness:** The degree of accuracy of annotations such as bounding boxes in relation to the actual objects and scenes collected by the sensor data.
- **Consistency:**Annotation uniformity across many frames and sensor modalities such as image-lidar alignment. Consistency is important to ensure reliable model training across time-series data.
- **Completeness:** Evaluation of whether all relevant data points and annotations are included. Missing labels or sensor readings reduce the model's performance.
- **Redundancy:** The identification of duplicate or unnecessary data, which introduces noise to training and testing processes, skewing results.

3.3 Detection and Validation Techniques

There are two types of detection techniques which are manual inspections and automated methods. Manual validation verifies the accuracy of annotations by visually comparing sensor data with the given labels, ensuring alignment between picture data and lidar point clouds. Automated detection scripts are used to measure error rates, get missing data, and detect duplicated entries. Consistency between variuos data modalities is also validated by comparing timestamps, alignment between image frames and corresponding lidar scans.

The outcomes of these studies are recorded and categorized according to the defined metrics, offering a thorough understanding of view of the data quality in both the KITTI and Waymo datasets.

4 Data Collection and Cleaning Plan

The process of data cleaning includes standardizing formats across the selected datasets, eliminating missing or incorrect information between sensor data and annotations. Any missing data points found are handled using advanced interpolation techniques. The cleaning process also has the removal of redundant data points that don't improve model performance.

Various imputation methods, such as k-nearest neighbors (KNN) and statistical approaches, are used to fill in the gaps in the datasets. The datasets are ready to be used after this stage.

5 Experiment and Data Analysis Plan

The experiment aims to assess how the identified problems can be mitigated for model performance, focusing on correctness, consistency, completeness, and redundancy. The following steps outline the experimental and analytical approach:

Phase 1: Quantitative Assessment. In the first phase, error rates and the percentage of missing data are calculated for both the KITTI and Waymo datasets. The percentage of redundant or superfluous data in the dataset is calculated to determine the proportion of duplicated or unnecessary data in the dataset. Statistical techniques like precision-recall and F1 scores are used to assess the correctness of labels.

Phase 2: Model-Based Impact Evaluation. This phase assesses training object detection models on both cleaned and uncleaned datasets to get the effect of data quality on performance. Models are computed in terms of accuracy, precision, recall, and computational efficiency. The performance of the models trained on the cleaned datasets is contrasted with models trained on the raw datasets to quantify the differences brought about by the data cleaning process.

Phase 3: Consistency and Completeness Analysis. The consistency of annotations across various time-series data is studied by examining the alignment between different sensor modalitie. Completeness is evaluated by finding instances where labels or sensor readings are missing, and their effect on model performance is calculated. The percentage of missing annotations is measured and compared with model degradation, providing details into the importance of comprehensive labeling.

Phase 4: Redundancy Removal Impact. Finally, the effect of redundancy removal on model performance is examined by comparing training times and accuracy before and after removal of redundant data points. Models trained on redundant data usually perform slower and may exhibit less accuracy due to overfitting.

This multi-phase experimental analysis gives a detailed understanding of how different data quality problem affect AV model performance. The results give valuable insights into improving the quality of AV datasets and give contribution to developing more reliable and dependant autonomous systems.

6 Task Assignment and Timeline

The project tasks are allocated among the members of the team in the following way:

- **Week 1 (Oct 11-17):** Selection of Dataset and Exploration (All members).
- **Week 2 (Oct 18-24):** Defining data quality metrics and developing scripts for data validation (Aiswarya Goriparthi, Nitya Vattam).
- **Week 3 (Oct 25-31):** Applying data quality metrics to the datasets selected (Vinuthna Reddy, Shanmukha varma).
- **Week 4 (Nov 1-7):** Comparing different data quality issues over KITTI and Waymo (Bhavya Madhuri, Pavani Kommineni).
- **Week 5 (Nov 8-14):** Focusing on specific data quality issues and conducting an analysis deeply (All members).
- **Week 6 (Nov 15-21):** Developing a framework to improve dataset quality (Aiswarya Goriparthi, Vinuthna Reddy, Bhavya Madhuri).
- **Week 7 (Nov 22-28):** Visualize results and preparing figures (Nitya Vattam, Shanmukha varma, Pavani Kommineni).
- **Week 8 (Nov 29-Dec 1):** Preparing final presentation (All members).
- **Week 9 (Dec 2-5):** Submitting the final report (All members).

References

[1] J. Smith and J. Doe, *Improving Sensor Calibration in Autonomous Vehicle Datasets*, IEEE Transactions on Intelligent Transportation Systems, vol. 22, no. 4, pp. 1234–1245, 2021.

[2] H. Liu and Y. Zhao, *Correcting Labeling Inaccuracies in Autonomous Driving Datasets Using Deep Learning*, Journal of Autonomous Systems, vol. 10, no. 2, pp. 345–356, 2020.

[3] S. Brown and J. Kim, *Ensuring Consistency in Multi-modal Autonomous Driving Datasets*, Autonomous Driving Journal, vol. 9, pp. 45–60, 2022.

[4] M. Perez and P. Garcia, *Redundancy Detection in Autonomous Driving Datasets*, Machine Learning for Transportation Systems, vol. 3, no. 1, pp. 88–95, 2019.

[5] X. Wang and Z. Li, *A Framework for Assessing Data Completeness in Autonomous Driving Datasets*, International Journal of Autonomous Vehicles, vol. 6, no. 4, pp. 312–325, 2021.

[6] M. Garcia and F. Rodriguez, *Detecting Label Errors in Autonomous Driving Datasets Using Statistical Techniques*, Autonomous Vehicles and Systems Journal, vol. 7, pp. 101–115, 2020.

[7] A. Martinez and L. Chen, *The Impact of Incorrect Annotations on Object Detection Performance in Autonomous Vehicles*, IEEE Transactions on Autonomous Systems, vol. 12, no. 1, pp. 67–78, 2023.

[8] K. Park and J. Lee, *Challenges in Ensuring Consistency Across Time-Series Data in Autonomous Driving*, Transportation Systems Journal, vol. 14, pp. 123–135, 2019.

[9] T. Henderson and D. Cooper, *Data Deduplication in Large-Scale Autonomous Driving Datasets*, Journal of Data Science for Transportation, vol. 5, pp. 145–158, 2022.

[10] P. Nguyen and K. Tran, *Automatic Detection of Errors in 3D Bounding Boxes in AV Datasets*, IEEE Transactions on Robotics and Automation, vol. 9, pp. 201–214, 2023.