AI-Powered Distance Estimation for Autonomous Systems: A Monocular Vision Approach

Sourabh Rajput
sourabh.rajput9232@gmail.com

***Abstract*—Accurate distance estimation is crucial for autonomous systems to navigate safely in dynamic environments. This study explores a novel deep learning-based approach to estimate object distances using monocular vision, eliminating the need for costly LiDAR sensors. By leveraging supervised neural networks and representation-based regression models, I improve real-time depth perception for self-driving applications. The methodology includes model adaptation, prototype development, and comparative performance evaluation across multiple architectures. Experimental results demonstrate the feasibility of low-cost, high-accuracy distance prediction, paving the way for enhanced situational awareness in autonomous vehicles.**

I. Introduction

Consider a scenario where an autonomous vehicle is navigating a roadway flanked by several other vehicles. To maintain a safe and adaptive cruising velocity, it must accurately gauge the proximity of the vehicle directly ahead. This research explores a cutting-edge methodology designed to infer the spatial gap between a camera and an object. The exploration encompasses four core phases: extensive literature review, replication of experimental outcomes from a seminal publication, architectural adaptation of the original model, and further refinement through novel modifications.

A. Motivation

Conventional methods estimate an object's distance using the following expression:

$$D = F_{\text{lens}} \times \frac{S_{\text{actual}}}{S_{\text{image}}} \tag{1}$$

However, practical deployments frequently lack knowledge of the real-world size of the object. In such situations, modern autonomous systems often rely on LiDAR, which utilizes light reflection delay to determine depth. While LiDAR ensures high precision, it is cost-intensive, typically ranging from $500 to over $8000 per unit. In contrast, monocular imaging systems offer a cost-effective substitute—priced around $125 per device—along with advantages such as resilience in adverse weather, chromatic information capture, and straightforward hardware integration.

Recent advances have shifted toward utilizing supervised deep learning architectures with monocular input for distance prediction. The process typically includes feature extraction, object identification, and training of neural regressors to predict either depth maps or absolute distances. Some implementations also integrate reinforcement learning techniques. Key obstacles in this domain include a lack of object class information, the absence of contextual scene understanding, reliance on single-frame RGB input, and the necessity for inference models to operate with minimal computational latency—suitable for real-time vehicular deployment.

II. Research Papers

Two key monocular vision methods were selected for implementation and extension after surveying state-of-the-art approaches.

A. Paper I: Object-Specific Range Prediction

[1] introduced the first deep network to directly predict object-centric distances from single-view images. The framework combines a CNN backbone, ROI pooling, and regression/classification heads for distance prediction and object categorization (Figure **??**).

B. Paper II: Representation Learning for Depth Regression

[2] adapted Convolutional Support Estimator Networks (CSEN) for regression tasks. The pipeline extracts object embeddings, constructs a dictionary $\mathcal{D}$, and approximates observations via collaborative representation (Figure **??**).

III. Other Approaches

Additional methods were reviewed but not implemented:

- **Relation Networks**: Effective for qualitative reasoning but unsuitable for continuous regression.
- **Graph Convolutional Models**: Classification-oriented with limited geometric estimation precision.

IV. Model Enhancement

The CSEN framework was validated using KITTI 3D dataset. Multiple CNN backbones were evaluated (Table **??**), with Xception demonstrating superior performance in consistency and error metrics.

V. Proximity Estimation Between Scene Entities

An auxiliary task was designed to estimate spatial gaps between objects in single frames:

- **Geometric method**: Predict depths and coordinates, then compute Euclidean distance
- **Learning-based method**: Directly infer inter-object distances from visual embeddings

The study focuses on the data-driven approach using the CSEN framework with KITTI dataset. Annotations were enriched with object IDs and pairwise distances (2D/3D Euclidean) for training the distance estimator.

VI. Implementation

A. Data Processing

Data was transformed for neural network compatibility using customized Python (feature extraction) and MATLAB (processing) scripts. Features were serialized into `.mat` files. From 140,000 object pairs, a balanced subset of 20,000 entries was selected to optimize computational efficiency while maintaining data diversity. The dataset was partitioned into five training runs of 100 epochs each.

B. Prototype

A baseline model was developed to estimate distances between object pairs:

1) Objects localized via KITTI bounding box annotations
2) Objects from same frame paired and individually cropped
3) Feature vectors extracted via backbone networks and concatenated
4) Ground truth distances used for supervised learning

C. Revised Architecture

An alternative approach used joint bounding boxes encompassing both objects to preserve spatial context. This composite region was processed by the backbone network while maintaining the same inference pipeline and supervisory signals.

D. Quantitative Assessment

Performance evaluated using:

- Mean Absolute Relative Discrepancy (ARD)
- Mean Squared Relative Discrepancy (SRD)
- Root Mean Squared Error (RMSE)
- RMSE Logarithmic

Metrics computed across five training instances (100 epochs each) with median and standard deviation reported.

1) Prototype Efficiency: Initial implementation showed RMSE of 9-10m for CSEN and 7-8m for CL-CSEN (Table **??**). CL-CSEN demonstrated slightly better precision, though significant inconsistencies were observed for short-distance predictions, potentially due to occluded objects or single-target design limitations.

2) Altered Configuration Performance: Comparative analysis (Tables **??** and **??**) revealed:

- Standard configuration generally outperformed altered variant
- ResNet50 most consistent in altered configuration
- Performance fluctuations likely due to random sampling from 140,000-record dataset

Both strategies demonstrated comparable effectiveness overall.

E. Experimental Evaluation

To explore the influence of dictionary dimensionality on model accuracy, a series of controlled experiments were performed using the CSEN architecture. Dictionary size, in this context, is determined by two key factors: the total number of support sets and the count of instances within each set. Under the existing scheme, sets consist of elements differing by exactly 1 meter in distance. When this constraint is relaxed—for example, allowing a 2-meter deviation—the number of distinct support sets is proportionally reduced. Experiments were carried out for support sets formed under tolerances of 1, 2, 5, and 7.5 meters, while simultaneously varying the number of internal samples across values of 10, 15, 20, and 25. Performance across these configurations is visualized in Figure 1 (prototype) and Figure 2 (altered prototype), evaluated using the Mean Absolute Relative metric.

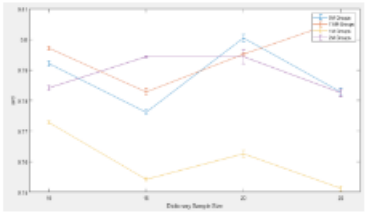

Fig. 1. Dictionary size variation: Original configuration

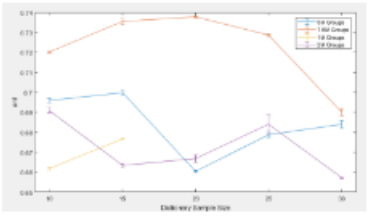

Fig. 2. Dictionary size variation: Modified configuration

The plots indicate that the restructured model demonstrates slightly improved accuracy when gauged using ARD. Furthermore, increasing the count of support sets contributes more significantly to accuracy than increasing the number of internal examples. For instance, comparisons between 2-meter and 7.5-meter groupings reveal higher reliability in denser set configurations. There is also an observable, albeit marginal, trend favoring higher intra-set sample counts.

In essence, while dictionary granularity affects predictive reliability, the total number of support groups exerts a more pronounced effect compared to individual sample volume per group.

VII. Conclusion and Future Work

A. Final Observations

The developed distance estimation framework, based on the CSEN-regressor architecture, successfully achieves its pri-

mary objective of delivering consistent and reasonably accurate inter-object distance predictions from monocular images. Through extensive experimental evaluation, we have gained significant insights into the model's operational characteristics, strengths, and limitations.

A key finding from this research is the demonstrated superiority of regression-based approaches over classification-oriented methods for continuous variable prediction tasks such as distance estimation. The CSEN and CL-CSEN frameworks have proven particularly effective in handling the continuous nature of distance regression, with CL-CSEN generally outperforming the basic CSEN implementation across multiple backbone architectures.

B. Limitations and Challenges

Despite achieving satisfactory performance, several limitations were identified during the evaluation:

- **Short-distance inconsistencies**: The model exhibits significant prediction variances for objects in close proximity, potentially due to occlusion effects or the inherent design bias toward single-object distance estimation in the original CSEN framework.
- **Sample sensitivity**: Performance fluctuations observed between different data subsets highlight the model's sensitivity to training data selection, emphasizing the need for robust sampling strategies.
- **Architectural constraints**: The requirement to maintain compatibility with the original CSEN structure limited certain architectural modifications that could potentially enhance performance.

C. Future Research Directions

Building upon the foundations established in this work, several promising avenues for future investigation emerge:

1) Angular Relationship Integration: A particularly compelling enhancement involves incorporating angular relationships between object pairs, which are readily available in the KITTI dataset annotations. This spatial information could be leveraged in multiple ways:

- **Independent distance proxy**: Angular relationships could serve as a complementary distance estimation mechanism, operating in parallel with visual feature-based approaches.
- **Feature fusion**: Angular data could be integrated with visual embeddings to create a richer, multi-modal input representation, potentially capturing more complex spatial dependencies.
- **Geometric constraints**: Angular information could enforce geometric consistency in predictions, reducing physically implausible outputs.

2) Architectural Innovations:

- **Attention mechanisms**: Incorporating spatial and channel attention modules could enhance the model's ability to focus on relevant visual cues for distance estimation.
- **Multi-scale feature fusion**: Leveraging features from multiple network layers could capture both local details and global contextual information.
- **Graph neural networks**: Modeling object relationships explicitly through graph structures could improve reasoning about spatial arrangements.

3) Data and Training Enhancements:

- **Semi-supervised learning**: Leveraging unlabeled data through consistency regularization or self-training approaches.
- **Curriculum learning**: Gradually increasing task difficulty during training to improve model robustness.
- **Multi-task learning**: Jointly learning related tasks such as object detection and depth estimation to improve feature representations.

4) Evaluation Framework Expansion:

- **Real-world validation**: Testing the framework on diverse real-world scenarios beyond the KITTI dataset.
- **Robustness analysis**: Systematic evaluation of performance under varying environmental conditions and object configurations.
- **Computational efficiency**: Optimization for deployment in resource-constrained environments such as autonomous vehicles.

The integration of angular relationships, in particular, represents a natural extension of this work that aligns with the spatial nature of the distance estimation task while maintaining connections to the original CSEN framework. This direction promises to enhance both the accuracy and robustness of monocular distance estimation systems for practical applications in autonomous driving and surveillance.

REFERENCES

[1] Jing Zhu et al. *Learning Object-specific Distance from a Monocular Image*. 2019. arXiv: 1909.04182 [cs.CV].

[2] Mete Ahishali et al. *Representation Based Regression for Object Distance Estimation*. 2021. arXiv: 2106.14208 [cs.CV].

[3] Andreas Geiger, Philip Lenz, and Raquel Urtasun. "Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite". In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2012.

[4] François Chollet. "Xception: Deep Learning with Depthwise Separable Convolutions". In: *CoRR* abs/1610.02357 (2016). arXiv: 1610.02357. URL: http://arxiv.org/abs/1610.02357.

[5] Christian Szegedy et al. "Rethinking the Inception Architecture for Computer Vision". In: *CoRR* abs/1512.00567 (2015). arXiv: 1512.00567. URL: http://arxiv.org/abs/1512.00567.

[6] Adam Santoro et al. *A simple neural network module for relational reasoning*. 2017. arXiv: 1706.01427 [cs.CL].

[7] Xiaolong Wang and Abhinav Gupta. *Videos as Space-Time Region Graphs*. 2018. arXiv: 1806.01810 [cs.CV].

[8] Mehmet yamaç et al. *Convolutional Sparse Support Estimator Network (CSEN) From energy efficient support estimation to learning-aided Compressive Sensing*. Mar. 2020.

[9] Kaiming He et al. "Deep Residual Learning for Image Recognition". In: *CoRR* abs/1512.03385 (2015). arXiv: 1512.03385. URL: http://arxiv.org/abs/1512.03385.

[10] Gao Huang, Zhuang Liu, and Kilian Q. Weinberger. "Densely Connected Convolutional Networks". In: *CoRR* abs/1608.06993 (2016). arXiv: 1608.06993. URL: http://arxiv.org/abs/1608.06993.

[11] Karen Simonyan and Andrew Zisserman. "Very Deep Convolutional Networks for Large-Scale Image Recognition". In: *CoRR* abs/1409.1556 (2015).

APPENDIX

Authors and Contributors

- **Carla**: Research Papers, Model Modification, Prototype
- **Tim**: Data acquisition & processing, Prototype & Performance(s), Conclusion
- **Benedikt**: Data Processing, Prototype, Altered Prototype, Experiments
- **Nils**: Introduction, Research Papers, Evaluating other Appraches, Model Modification