Anteproyecto de Trabajo Fin de Grado

Esteban Gómez García

4 de octubre de 2024

Autor: Esteban Gómez García

Tutores: Luis Miguel Bergasa Pascual y Santiago Montiel Marín

Titulación: Grado en Ingeniería Informática

Título: Deteccion de objetos 3D con metodos deep learning y fusión temporal de cámara y LiDAR

Departamento: Departamento de Electrónica

1 Introducción

En los últimos años, la detección de objetos en tres dimensiones (3D) ha adquirido un papel crucial en el avance de los sistemas de conducción autónoma, representando un desafío significativo para garantizar la precisión y seguridad en la interpretación del entorno. La combinación de sensores LiDAR y cámaras permite capturar tanto información de profundidad como de color y textura, aspectos vitales para una comprensión integral del entorno en el que operan estos sistemas. Sin embargo, los métodos tradicionales que procesan datos cuadro por cuadro presentan limitaciones, como la sensibilidad al ruido, oclusiones y la dispersión de datos, lo que puede afectar la robustez y consistencia de la detección.

Este anteproyecto explora la integración de técnicas de Deep Learning con métodos de fusión temporal que emplean múltiples entradas de sensores, como cámaras y LiDAR, para mejorar la detección de objetos en 3D. Se propone la implementación de un modelo que, inspirado en la arquitectura de detección de objetos Temp-Frustum Net, fusione características de cuadros anteriores mediante un Módulo de Fusión Temporal (TFM). Este módulo permite combinar la información de diferentes instantes de tiempo para compensar las limitaciones de los modelos basados en cuadros individuales, incrementando así la robustez y precisión de la detección ante oclusiones y situaciones complejas de tráfico.

El desarrollo de este sistema se validará utilizando el dataset KITTI, ampliamente reconocido en la investigación de conducción autónoma, y se implementará en dispositivos NVIDIA Jetson para evaluar su viabilidad en entornos hardware de bajo consumo.

2 Objetivos

El objetivo fundamental de este proyecto es el desarrollo e implementación de un sistema de detección y estimación de objetos en 3D con métodos de Deep Learning que combinen modelos basados en redes neuronales y técnicas de fusión temporal para datos de cámara y LiDAR, validando su desempeño tanto en entornos simulados como en condiciones reales.

Los objetivos específicos de este proyecto son los siguientes:

- Estudio y análisis de la arquitectura YOLO [1] y de la red propuesta en el modelo Temp-Frustum Net [2]:
  1. Realizar una revisión exhaustiva de la arquitectura YOLO, destacando sus capacidades y limitaciones en la detección y estimación de objetos 3D a partir de imágenes 2D. Analizar el funcionamiento de la arquitectura Temp-Frustum Net y sus innovaciones en la integración de características temporales mediante el Módulo de Fusión Temporal (TFM), comprendiendo su aplicación en la detección de objetos en 3D.
- Análisis y comprensión del dataset KITTI:
  1. Estudiar la estructura, características y contenido del dataset KITTI [3], evaluando su utilidad para entrenar y validar modelos de detección de objetos en 3D. Adaptar y preparar los datos del dataset KITTI para su uso eficiente en los modelos de detección, garantizando una segmentación precisa.
- Implementación y fine-tuning de la arquitectura YOLO para detección 3D:
  1. Implementar y adaptar la arquitectura YOLO para su entrenamiento en el dataset KITTI, ajustando sus parámetros y estructura para la estimación tridimensional de objetos.
- Desarrollo del Módulo de Fusión Temporal (TFM) y su implementación en el sistema:
  1. Integrar el Módulo de Fusión Temporal (TFM) para mejorar la detección 3D mediante el uso de características temporales que provienen de cuadros sucesivos, incrementando la precisión ante oclusiones y situaciones complejas de tráfico con la fusión de datos LiDAR y de cámara.
- Validación del sistema en el dataset KITTI:
  1. Entrenar y evaluar los modelos implementados utilizando el dataset KITTI, midiendo su precisión y eficiencia en la detección de vehículos, peatones y otros objetos en condiciones de tráfico diversas.
- Optimización, mejoras del rendimiento y despliegue:
  1. Implementar estrategias de optimización para reducir los tiempos de inferencia y mejorar la precisión de detección.
  2. Despliegue en el dispositivo NVIDIA Jetson.

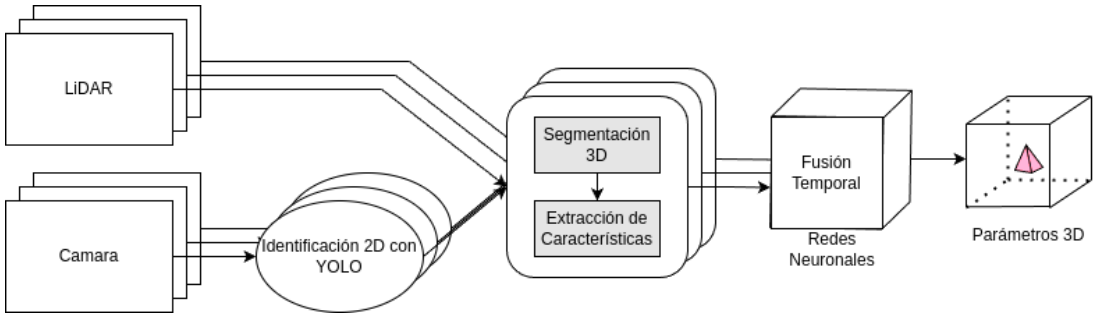

Figura 2: Flujo del sistema propuesto.

3 Metodología y plan de trabajo

El proyecto a realizar tiene un duración aproximada de 8 meses, entre noviembre de 2024 y junio de 2025. El desarrollo se divide en fases para el cumplimiento de los objetivos del proyecto descritos en la sección 2:

1. Formación inicial y revisión del estado del arte (1 mes)
    - Estudio de la arquitectura YOLO.
    - Estudio del framework para machine learning PyTorch.
    - Consulta bibliográfica de modelos de deteccion de objetos con deep learning y fusion temporal de datos a partir de camaras y LiDAR.
2. Estudio detallado de la estructura del dataset KITTI (0,5 meses)
    - Comprension de la organizacion de los datos y objetos incluidos.
    - Estudio de las matrices de calibración pertenecientes a las cámaras.
    - Estudio de las nubes de puntos LiDAR.
3. Implementacion y ajuste de la arquitectura YOLO para la deteccion 3D (2 meses)
    - Adaptacion de YOLO para la deteccion 3D.
    - Realizar fine-tuning para mejorar la precisión.
4. Desarrollo del módulo de Fusión Temporal (TFM) y su integración en el sistema (2 meses):
    - Integrar técnicas de fusión temporal para aprovechar multiples cuadros de datos.
5. Evaluación y validación del sistema en el dataset KITTI (1 mes)
    - Pruebas de validación y analisis de métricas.
6. Optimización y mejora del rendimiento (0,5 meses)
    - Aplicación de técnicas de optimización.
    - Despliegue en el dispositivo NVIDIA Jetson.

7. Documentación y escritura del informe final (1 mes)
    - Documentar el proceso, estudios y resultados.
    - Redacción de la memoria final y presentación de resultados.

4 Medios

Las herramientas necesarias para desarrollar este proyecto de forma correcta son las siguientes:

- Componentes Hardware:
    - PC con memoria RAM de 32 GB y tarjeta grafica NVIDIA RTX 3060.
    - Sensores LiDAR y cámaras de alta resolución.
    - Dispositivo NVIDIA Jetson.
- Componentes Software:
    - Sistema operativo Ubuntu 22.04 LTS.
    - Lenguaje de programación Python.
    - Frameworks de Deep Learning: PyTorch [4]
    - Procesador de textos LaTeX **p**ara la documentación del proyecto.
    - Herramientas de control de versiones Git.
    - Acceso a datasets como KITTI.

Referencias

[1] J. Redmon, S. Divvala, R. Girshick y A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection", en *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Disponible en: https://arxiv.org/abs/1506.02640, 2016.

[2] E. Ercelik, E. Yurtsever y A. Knoll, "Temp-Frustum Net: 3D Object Detection with Temporal Fusion", en *2021 IEEE Intelligent Vehicles Symposium (IV)*, Nagoya, Japan: IEEE, jul. de 2021, págs. 1095-1101. DOI: 10.1109/IV48863.2021.9575392.

[3] A. Geiger, P. Lenz, C. Stiller y R. Urtasun, "Vision meets Robotics: The KITTI Dataset", en *Proceedings of the International Journal of Robotics Research (IJRR)*, 2013.

[4] A. Paszke, S. Gross, F. Massa et al., "PyTorch: An Imperative Style, High-Performance Deep Learning Library", en *Proceedings of the International Conference on Neural Information Processing Systems (NeurIPS)*, Disponible en: https://arxiv.org/abs/1912.01703, 2019.