NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Global Connect Fellowship

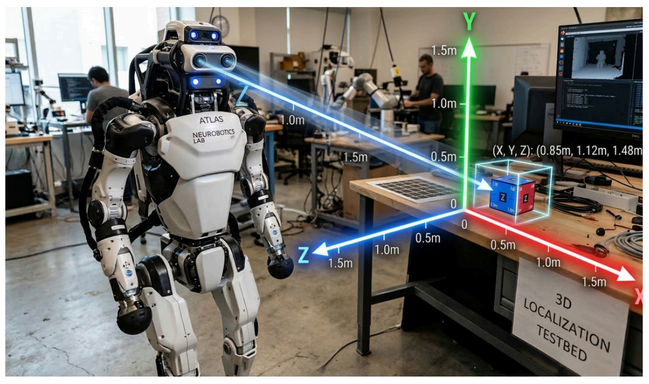
Lightweight Object Depth Estimation Using Restricted Coulomb Energy Neural Networks for Constrained Stereo Vision

Darius Dayu Giannoli, EPFL · Assoc. Prof. Xie Ming, Humanoid Robotics & AI Group, NTU

School of Mechanical and Aerospace Engineering, Nanyang Technological University

INTRODUCTION & OBJECTIVES

The Primacy of Depth: Visual signals are the foundation of humanoid intelligence. Depth perception is the "entry point" for the Perception →Planning → Control loop, enabling 3D navigation and manipulation, today's challenges.



Objectives: A Biomimetic Alternative

Current depth sensing (LiDAR, Radar) and Deep Learning are computationally expensive and lack transparency.

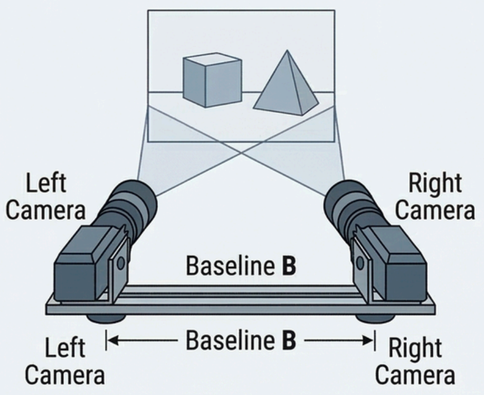
Our Goal: A human-like binocular stereo system using RCE Neural Networks to achieve:

- Low Cost: Real-time performance on standard edge-CPUs (no GPUs).
- High Speed: Sub-20ms latency for dynamic humanoid movement.
- XAI: Transparent, physically-grounded "DNA" matching logic allowing to manually debug and ensure accuracy

STEREO DEPTH PIPELINE OVERVIEW

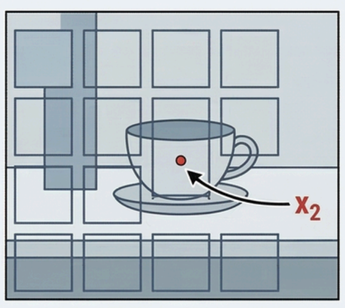
1. STEREO PAIR

Two calibrated cameras capture the same scene from slightly different viewpoints, separated by baseline B.



2. RECOGNITION & LOCALIZATION

A sliding window or feature detector scans the **Right Image** for **candidate patches**. Each patch is classified (e.g., by a CNN) to localize object features, defining position $\mathbf{x}_2$.

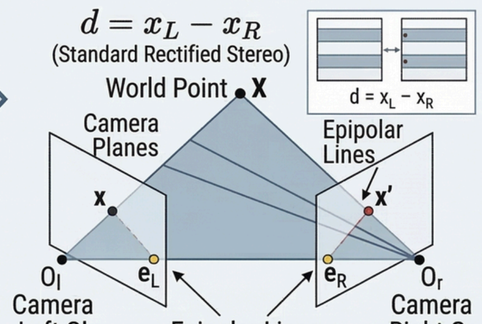


3. EPIPOLAR GEOMETRY

The epipolar constraint connects corresponding points across views. worispplane. Matching points must lie on specific **conjugate epipolar lines**, reducing the search space from 2D to 1D.

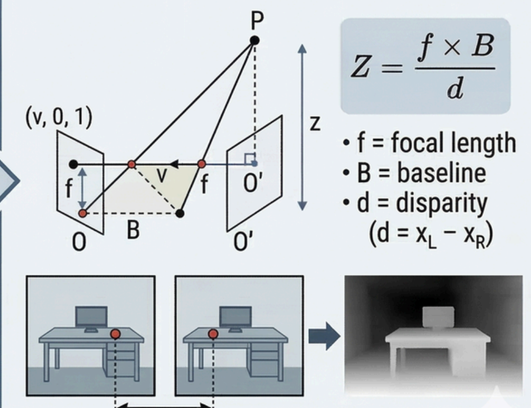
$$d = x_L - x_R$$

(Standard Rectified Stereo)



4. DISPARITY

Disparity (d) is **inversely proportional** to **depth** (Z). Closest objects have larger disparity.



$$Z = \frac{f \times B}{d}$$

- f = focal length
- B = baseline
- d = disparity ($d = x_L - x_R$)

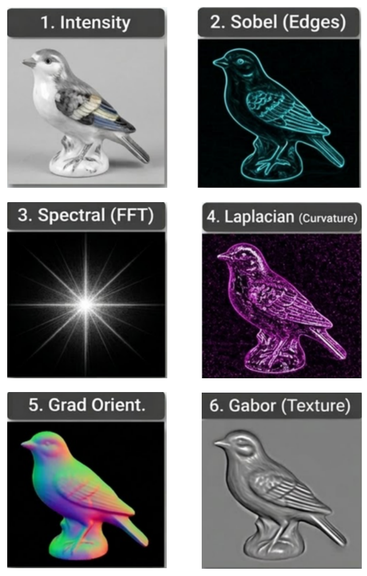
PROPOSED RECOGNTION SOLUTION & VALIDATION

Physics-Encoded Feature DNA: The Modular Adaptive Registry

Vector Fusion: 10-bin modular histograms are concatenated into a compact, adaptive 60D "DNA" feature vector.

$$\mathbf{f} = [H_1 \,\|\, H_2 \,\|\, \cdots \,\|\, H_N] \in \mathbb{R}^{N\times 10}, \quad N \in \{1, \ldots, 6\}$$

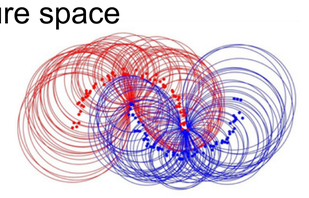
A unique physical fingerprint for object identity, enabling transparent region comparisons bypassing DL inference.



Module	Mathematical Descriptor (H_X)	Target Feature
Intensity	$H_I(b) = \sum_{x,y} \mathbf{1}[I(x,y) \in b]$	Global Brightness
Sobel	$H_S(b) = \sum_{x,y} \mathbf{1}\left[\sqrt{G_x^2 + G_y^2} \in b\right]$	Edge Strength
Laplacian	$H_L(b) = \sum_{x,y} \mathbf{1}\left[\lvert\nabla^2 I(x,y)\rvert \in b\right]$	Surface Curvature
Grad. Orient.	$H_O(b) = \sum_{x,y} w_{x,y}\, \mathbf{1}\left[\arctan\frac{G_y}{G_x} \in b\right]$	Edge Directionality
Spectral	$H_F(b) = \sum_{r\in b} \log(\lvert\mathcal{F}(I)\rvert + 1)$	Texture Periodicity
Gabor	$H_G(b) = \frac{1}{4}\sum_\theta \mathbf{1}[\lvert G_{\lambda,\theta} * I\rvert \in b]$	Oriented Texture

Restricted Coulomb Energy Neural Networks (RCE NN)

Physics-inspired engine for human-like cognition and incremental learning. It maps object signatures as dynamic hyper-spheres in the feature space

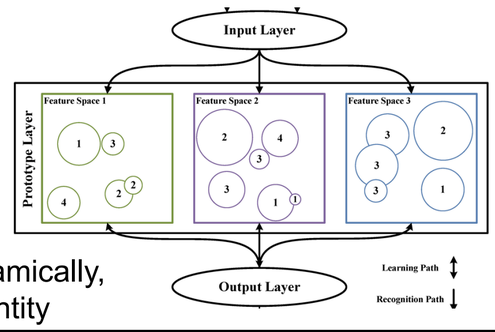


1. Cognition (Learning Path): Memorizes entities by defining hyper-spheres in feature space centered at the mean vector:

$$F_{i,a} = \frac{1}{k_i}\sum_{j=1}^{k_i} F_{i,j}$$

2. Recognition (Matching Path): Evaluates candidate matches via a Possibility Function to measure the "belief" in an identity:

$$p_{j,n}(i) = e^{\frac{-1}{\sigma_i^2}(F_{0,j}-F_{i,a})^T(F_{0,j}-F_{i,a})}$$



N-Node Growth: Prototype Layer expands dynamically, adding a hyper-sphere node for every learned entity

Perception Lab: Challenging Black-Box Vision with Recurrent Coulomb Energy NN

webapp:

A real-time framework for benchmarking few-shot recognition and 3D spatial awareness:

- Stereo Depth Analysis: Comparative evaluation of SOTA backbones, localization strategies, and the custom RCE network against traditional Depth NNs.
- Multi-View Generalization: Quantifying recognition robustness and cross-view consistency across calibrated stereo image pairs.
- On-the-Fly Adaptation: Instant model evolution via real-time "head" fine-tuning on user-provided or curated datasets
- Explainable AI (XAI): Moving beyond the "black box"—visualizing internal feature maps and energy landscapes

Architecture	Strategy	Technical Advantage
ResNet-18	Deep CNN	High-dimensional feature stability for complex textures.
MobileNetV3-S	Efficient CNN	Minimal parameter count; optimized for edge-AI latency.
MobileViT-XXS	Hybrid ViT/CNN	Global spatial context through multi-headed self-attention.
YOLOv8n	Regression	Direct bounding box prediction with one-stage inference.
ORB	Classical Geom.	Transparent, rotation-invariant keypoint descriptor matching.
RCE Network	Energy Minimization	**Proposed:** Vectorized, non-black-box spatial logic.

Method	Description & Algorithmic Logic
Sliding Window	Brute-force grid scan evaluating the recognition head at every pixel coordinate.
Image Pyramid	Multi-scale search processing image resizes to handle distance and scale variance.
Coarse-to-Fine	Two-pass hierarchical refinement; low-res candidate detection followed by high-res local search.
Contour Proposals	Edge-driven selection using Canny/Morphology to isolate candidates; 10–100× faster than exhaustive.
Template Matching	Fast baseline using OpenCV cross-correlation (NCC); runs in optimized C++ backend.

Acknowledgements

I express my sincere gratitude to **Professor Xie Ming**, **Nanyang Technological University (NTU)**, and the **GCF personnel** for their invaluable support and technical guidance during this internship.

Contact: darius.giannoli@epfl.ch