Vision LLMs in Multi-Agent Collaborative Systems: Architecture and Integration

Niyati Prajapati

ML/Gen AI Lead

Google

Google

Overview

Agenda

- Context
- Intersection of machine vision, large language models (LLMs) and human computer interface (HCI)
- Multi-agent collaborative systems (MACS) architecture, integration, performance
- Vision LLMs into MACS
- Real world use cases
  - Sensor fabrication analysis systems
  - Warehouse robotics systems
- Future trend and challenges

Context

Context

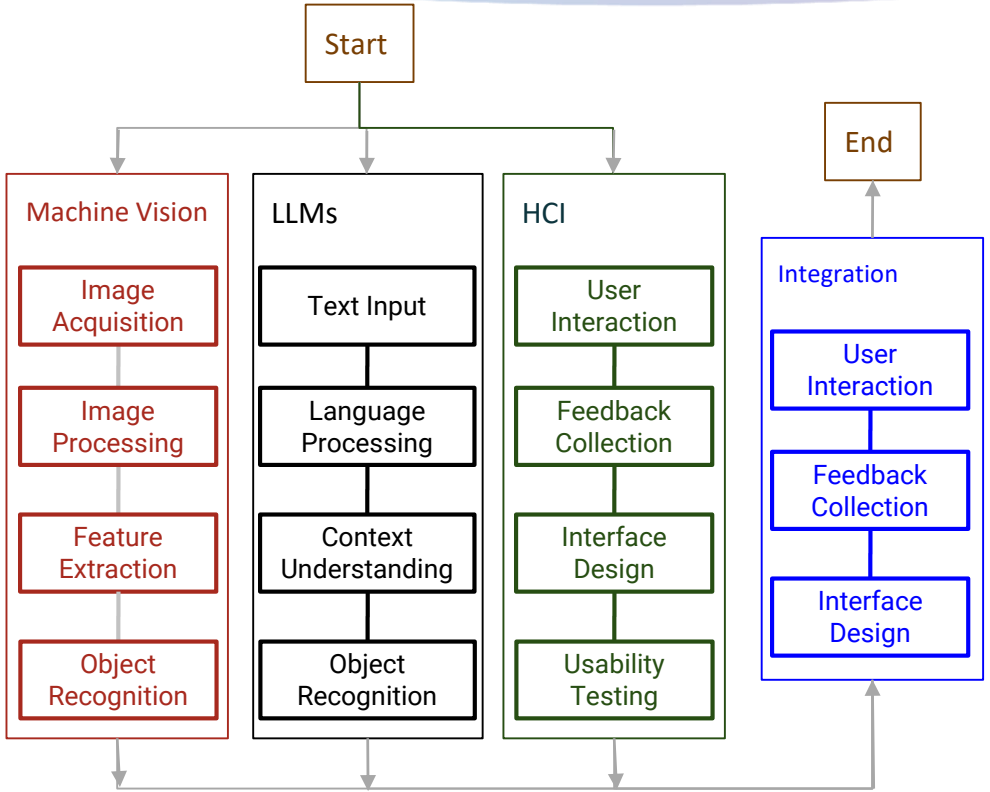
Urgency

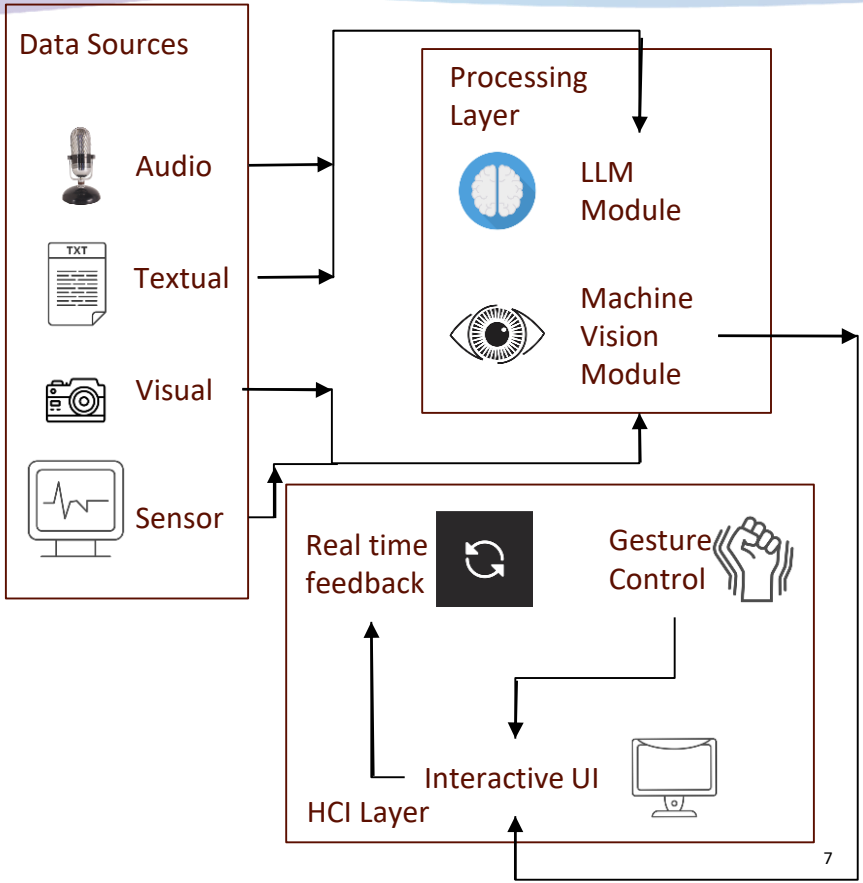
- Rising complexity of data, automation, robotics, logistics and manufacturing
- Increasing demand of autonomous decision making
- Limitation of legacy systems without advanced vision, reasoning and efficient collaborative capabilities

- What makes vision LLMs + multi-agent collaborative systems (MACS) ground breaking? Next-gen architecture with the fusion of visual reasoning and distributed intelligence.
- **“Vision LLMs = smarter cameras”** — understand scenes *semantically* using natural language
- MACS — allows distributed agents to reason collectively with intelligent perception and coordination across agents.
- Central outcome: “autonomous decision-making system”

Intersection of machine vision, LLMs and HCI

Intersection of machine vision, LLMs and HCI





Intersection of machine vision, LLMs and HCI

What makes vision LLMs + MACS synergy practical with HCI?

- Multimodal LLMs, real-time multi-agent frameworks, distributed systems that support these functionalities at scale
- Autonomous systems that perceive and interpret their environments to coordinate actions and interact with humans intuitively.
- Triad that forms the core of next-gen AI-driven industrial and human-interactive systems with below key technologies :
  - Vision-language models (GPT-4V, PaLI, Kosmos-2)
  - Reinforcement learning for agent coordination
  - Distributed multi-agent frameworks (ROS2, Ray)
  - Edge AI & vision hardware (Jetson, RealSense)

Multi-agent collaborative systems (MACS) architecture, integration and performance

Multi-Agents Collaborative Systems architecture, integration and performance

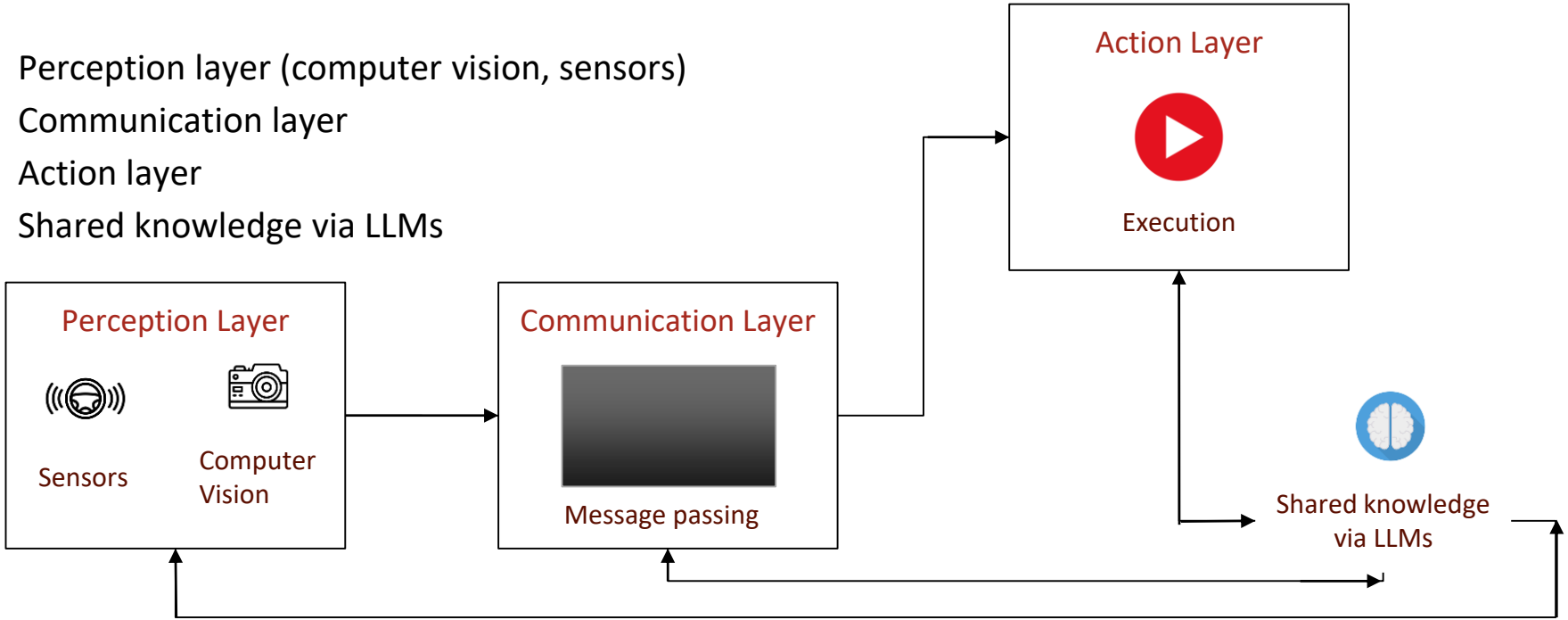
Core concepts

- Vision LLMs[1]
  - Inputs: Image + optional text
  - Outputs: Descriptive reasoning, decision support
  - Capabilities: Q&A, scene interpretation, captioning, semantic analysis
- MACs:
  - Collaborative distributed agents performing autonomous/semi-autonomous operation
  - Communication: Agents exchange visual, linguistic and task level signals
  - Perception: Each agent interprets local sensor data using Vision LLMs
  - Shared objectives: All agents operate toward a unified goal for adaptive decision-making protocols.

Multi-agent collaborative systems architecture, integration and performance

MACS Architecture

- Perception layer (computer vision, sensors)
- Communication layer
- Action layer
- Shared knowledge via LLMs



Multi-agent collaborative systems architecture, integration and performance

Vision LLMs into MACS

- Shared visual understanding through natural language
- Vision LLMs = perceptual interpreters and planners
- Decoupled sensing + reasoning

Performance Benefits

- Better perception in complex environment
- Scalable, robust and explainable

Multi-agent collaborative systems architecture, integration and performance

MACS with vs without Vision LLMs[2]

Metric	MACS without Vision LLMs	MACS with Vision LLMs	Improvement	Analogy
Package sorting accuracy	90%	98%	+8%	Specialized workers with fixed signals vs intelligent workers with language
Handling time per package	15 seconds	12 seconds	-20%	Step-by-step visual check vs contextual understanding
Adaptability to new package types	weeks	hours/minutes	significantly faster	Manual retraining for each type vs understanding through description
Communication efficiency	moderate	high	increased	Fixed codes vs natural conversation
Handling ambiguity	low	high	significantly Better	Relying solely on appearance vs using language for clarification

Real world use case 1:
Sensor fabrication analysis system

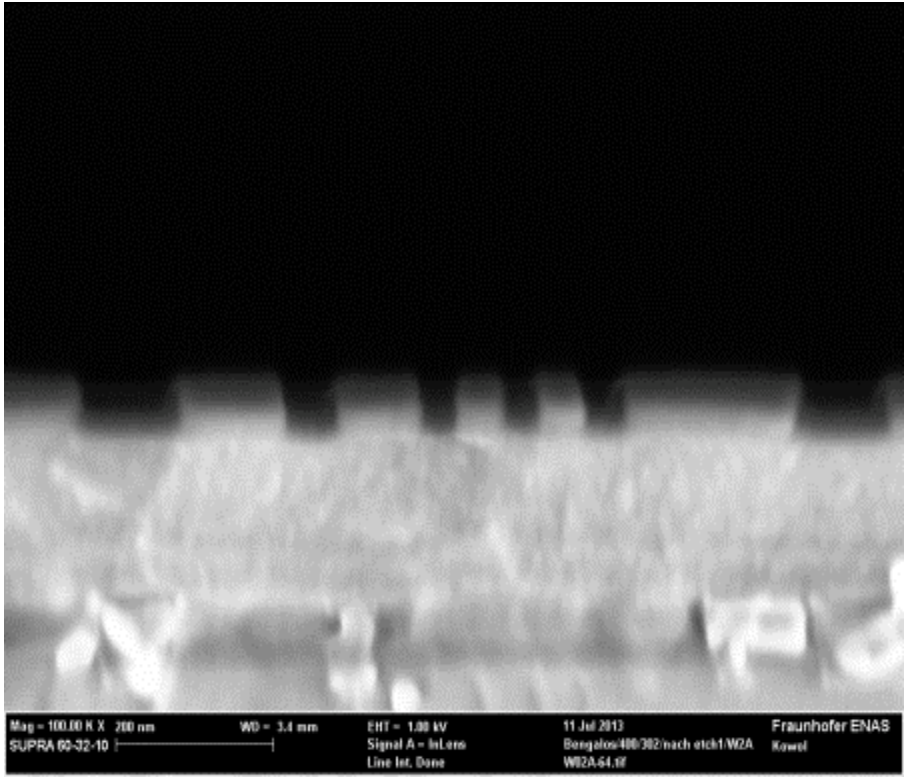
Real world use case - Sensor fabrication analysis system

MACS Architecture

- Scanning electron microscope (SEM) image of carbon nanotube (CNT) fabrication

  [Scanning / transmission electron microscopy]

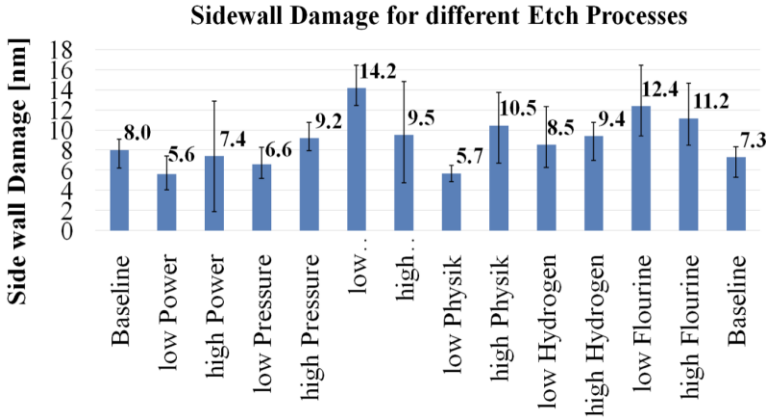
- Domain use cases:
  - Precision + automation in micro-nano manufacturing
  - High-resolution imagery + interpretation



Sensor fabrication analysis system

CASE A: Automated quality control and defect detection[3]

- **Agent1:** Acquires SEM/TEM images
- **Agent2:** LLM defects from etching and polymerization process results
- **Agent3:** Predicts performance quality with linear coefficient to measure identified species in plasma and different geometrical parameters like undercut, line edge roughness, sidewall, damage, linewidth
- **Outcome:** Real-time quality control and feedback



3 agents performing collective actions of acquiring image in analysis software, perform defect analysis, predict overall fabrication and performance quality with decision making

Sensor fabrication analysis system

CASE A: Automated quality control and defect detection[3]

Agents Functionalities:

- **Agent 1:** Captures high-resolution images from electron microscopes (SEM/TEM) and surface analysis instruments during CNT synthesis and deposition stages

  [Required pre-processing - denoising, contrast enhancement, edge detection]

- **Agent 2:** Detects and categorizes CNT imperfections [i.e. misaligned bundles, structural defects]

Vision-Language Interpretation:

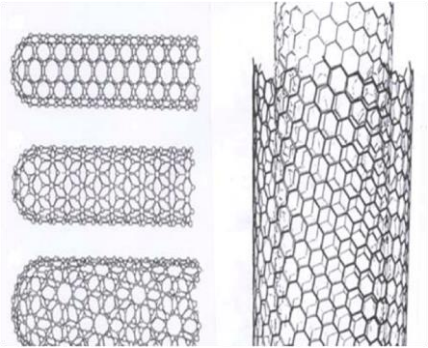
- Describes & generates visual defects in natural language terms (“Surface irregularity , misalignment of CNT bundles”)

- **Agent 3:** Quality performance prediction Integrating data from fabrication, visual scans, and defect reports
- Triggers corrective actions (e.g., adjusting chemical vapor deposition parameters).
- **Outcome:** Real-time quality control and feedback

Real world use case - Sensor fabrication analysis system

CASE B: Nano-scale pattern recognition for CNT structural analysis[4]

- **Agent1:** Captures structural imagery
- **Agent2:** LLM interprets morphology
- **Agent3:** Material classification and analysis
- **Outcome:** AI enhanced feedback loop for research and development



3 agents collects data of CNT imagery, provide categorical and structural interpretation about what kind of CNT is identified and perform material analysis with AI driven data pipelines

Sensor fabrication analysis system

CASE B: Nano-scale pattern recognition for CNT structural analysis[4]

Agents Functionalities:

- **Agent 1 :** Capturing high-resolution CNT structures using imaging tools like SEM or AFM
- **Agent 2 :** Vision LLM interprets and categorize CNT morphological patterns [e.g., alignment, defects,density]
- **Agent 3 :** Classifies the material type and predicts performance properties [e.g., conductivity, strength]
- **Continuous AI-driven pipeline**: sensing → interpretation → material analysis → feedback for **accelerating R&D cycles**

Real world use case 2: Warehouse robotics

Real world use case - Warehouse robotics

Introduction

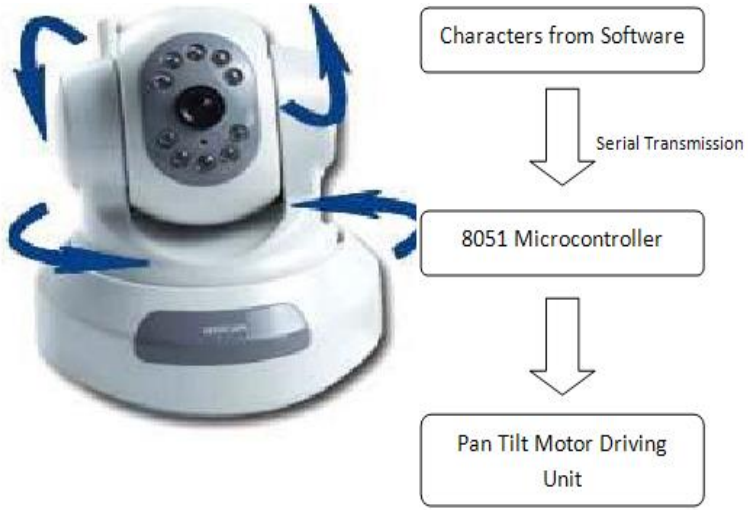
- Need visual intelligence and coordination in logistics
- Dynamic cluttered environments

Figure A: Robotics warehouse[5]

Real world use case - Warehouse robotics

CASE A: Inventory and tracking

- **Agent1:** Pan-tilt zoom camera scan
- **Agent2:** Vision LLM analyzes and learns inventory status
- **Agent3:** Robot fleet coordination
- **Outcome:** Efficient pathing, fewer errors



Inventory tracking, object recognition and dynamic path optimization: multi agent system workflow

Agents workflow: multi camera scanning, visual reasoning through interpretation and adaptive learning by coordinating robot fleet system

Real world use case - Warehouse robotics

CASE A: Inventory and tracking

Agents Functionalities :

- **Agent 1 :** Continuous wide-area pan-tilt-zoom (PTZ) camera scanning to to capture and update the warehouse's visual inventory
- **Agent 2 :** Vision LLM to interpret visual data, inventory items, detect misplaced or missing goods, and update the inventory
- **Agent 3 :** Manages the robot fleet’s routes for picking, stocking, and repositioning items
- **Vision Acquisition Agent → Perception and Reasoning Agent → Action Coordination Agent**

Real world use case - Warehouse robotics

CASE B: Real-time Fault Detection[5]

- **Agent1:** Detects faults visually
- **Agent2:** Root cause reasoning
- **Agent3:** Dispatch + system update
- **Outcome:** Downtime reduction, proactive maintenance

Real-time fault detection and maintenance: multi-agent system workflow

Agents workflow: Vision fault detection, root cause analysis and real-time update on maintenance status

Real world use case - Warehouse robotics

CASE B: Real-time Fault Detection[5]

Agents Functionalities :

- **Agent 1 :** Continuously monitors machinery, conveyor belts and robotic systems through visual data to detect anomalies
- **Agent 2 :** Vision LLM reasoning to diagnose the underlying cause of detected misalignments, mechanical wear
- **Agent 3 :** Dispatch of maintenance bots and updates system logs to prioritize and track repair tasks
- **Fault Detection Agent → Root Cause Analysis Agent → Maintenance Coordination Agent**

Future trends and challenges

Future trends and challenges

Future Trends

- Vision language graph agents
- Edge deployment of LLMs
- Neuro-symbolic MACS
- Closed loop learning

Challenges

- Real-time performance
- Agent safety and control
- Communication protocol standardization

Conclusion

Case	Extended applied	Improvements quantified
CNT fabrication - quality control	Pilot/Research labs	15–25% defect detection ↑, 2–3x inspection speed ↑, 8–12% yield ↑
CNT fabrication - pattern recognition	R&D labs	20% structural ID ↑, 30–40% cycle time ↓
Warehouse - inventory and tracking	Deployed	35% inventory error ↓, 20–30% efficiency ↑, 15–25% order speed ↑
Warehouse - fault detection	Deployed/Vision LLM	50–60% faster fault detection, 20–40% downtime ↓

- Vision LLM + MACS = future of intelligent automation
- CNT fabrication and warehouse logistics as real-world applications
- Synergy of vision, language and collaboration
- Vision LLM improves interpretability for human operators

References

References

[1] PaLI: A Jointly-Scaled Multilingual Language-Image Model, Chen, X., Alayrac, J.-B., *Google Research* et al. (2023).

[2] Multimodal capabilities of GPT-4V for real-world vision-language tasks, OpenAI, GPT-4 Technical Report, 2023.

[3] Automated SEM/TEM Defect Analysis in CNT Electronics Fabrication, Samsung Research Whitepaper, Nanoelectronics and Materials Division, 2022.

[4] Automated Characterization of Carbon Nanotube Arrays Using Computer Vision Technique, IEEE Trans. Nanotech., Vol 21, 2022, DOI: 10.1109/TNANO.2022.3149572

[5] Warehouse Robotics and Vision-based Automation for Inventory Management, Amazon Robotics 2022 Whitepaper, “Towards Fully Autonomous Fulfillment Centers”.

Question and discussion