Democratic and Popular Republic of Algeria

Ministry of Higher Education and Scientific Research

University Centre of Illizi

Institute of Science

Department of Computer Science

Artificial Intelligence & its Applications Master's Thesis

Rotative Robotic Arm for Intelligent Face Tracking

Authors:

Tagabou Doua

doua.tagabou@cuillizi.dz

Advisors:

Dr. Khelifi Ramzi

khelifi.ramzi@cuillizi.dz

Examining Committee:

- (President) Dr. Bezza Youssef
  bezza.youcef@cuillizi.dz
- (Examinator) Dr. Kahla Mohammed El Habib
  kahla.med-elhabib@cuillizi.dz

June 2026

Acknowledgments

All praise and gratitude belong to Almighty Allah, by whose grace and blessings this work has been successfully completed.

First and foremost, I would like to express my deepest gratitude and profound appreciation to my esteemed academic supervisor, Dr. Ramzi Khelifi, for his invaluable guidance, continuous support, and insightful remarks throughout the preparation of this thesis. I am deeply thankful for the extensive time and dedication he committed to reviewing this work with immense patience and meticulous attention, which significantly enhanced the quality of this research.

My sincere thanks also extend to all the professors who contributed to my academic foundation and academic journey throughout my university studies.

In a gesture of deep affection, I would like to express my special thanks to my beloved family and dear friends for their endless prayers, constant encouragement, and unwavering support during all stages of my education.

Furthermore, I am highly grateful to everyone who extended a helping hand, whether near or far, and to all those who supported and motivated me in accomplishing this humble piece of work.

Finally, I pray to Almighty Allah to accept this work as a sincere deed for His sake and to make it a source of benefit, knowledge, and inspiration for others.

DEDICATION

In the name of Allah, the Most Gracious, the Most Merciful.

All praise is due to Allah Almighty, by whose grace all good deeds are accomplished. He has granted me strength, patience, and determination to complete this humble work, marking the conclusion of an important stage in my academic journey. I sincerely pray that it may be beneficial knowledge and purely devoted to His noble countenance.

To my beloved mother, may Allah protect her and grant her health and long life.

You are my lifelong support, my safe haven, and the soul behind every success I have achieved. After Allah, nothing in my life would have been possible without your love, prayers, and sacrifices. To you, I dedicate this work with all my heart, praying that Allah rewards you abundantly and blesses you as you have always blessed my life.

To my dear brothers: Tagabou Islam, Tarek, Abdelkader, and Mohamed, who have been a constant source of support and strength, sharing both my hardships and moments of joy, and encouraging me throughout all stages of my academic life.

To my beloved fiancé, Mohammadi Daoud,

Thank you for standing by my side, for your continuous support and encouragement, and for every word of hope you offered whenever I needed strength and confidence to move forward.

To the cherished souls who have left this world, yet remain forever present in my heart and prayers...

To my late grandfather Boughrari Mohamed, my late grandmother Bousmaha Cheikha, my late cousin Zegri Mohamed Ramadan, and to all my beloved ones whom I have lost but whose memory continues to live within me. To your pure souls, I dedicate this work, and I ask Allah to shower you with His mercy and grant you the highest ranks of Paradise.

To my dear friends: Cherchef Akila, Soumaa Nardjes, and their families, thank you for every word of encouragement, every moment of support, and every beautiful memory that made this journey lighter and more meaningful.

To all members of my extended family, and to everyone who supported me, believed in me, encouraged me, or stood by me with a kind word, a prayer, or sincere help, whether near or far,

I dedicate this humble work to all of you, hoping it reflects even a small part of your love, support, and trust.

To all of you... I dedicate this achievement.

Abstract

TeleGimbal is an intelligent rotating robotic arm powered by artificial intelligence for face tracking. The system aims to solve the problem of target loss caused by the continuous movement of speakers inside lecture halls, where fixed cameras or human operators often fail to track movement reliably and cost-effectively.

The system relies on computer vision algorithms using the OpenCV library, where faces are detected using the YuNet model and recognized using the SFace model. Based on the visual coordinates obtained from image processing, the system sends real-time control commands through an Arduino UNO microcontroller to drive the Pan rotation axis and the Tilt axis using NEMA 17 stepper motors controlled by A4988 drivers with 1/16 microstepping enabled, providing smooth, silent, and vibration-free motion.

Five major engineering challenges were documented and addressed: audio resonance, solved through microstepping activation; overshoot caused by artificial intelligence processing, solved by applying a 50-pixel tolerance threshold and limiting speed; mechanical vibration, solved by physical adjustments and increasing the mounting height; target loss, solved through a safety mechanism that immediately sends a stop command; and stepper motor overheating, solved by adjusting the VREF value to 0.8 volts.

Experimental results demonstrated high operational efficiency, achieving frame rates ranging from 27.75 to 29.79 frames per second, face recognition confidence exceeding 0.7, response times below 50 milliseconds, and an angle measurement accuracy reaching 92.14 degrees. From a commercial perspective, the feasibility study showed a gross profit margin of 71.6%, a break-even point at 41 units sold, and a return on investment reaching 133.6% at 200 units sold.

This academic prototype was developed within the framework of Ministerial Decision No. 1275 and is proposed as a marketable commercial product under the startup project TeleGimbal.

Keywords: Computer Vision, Face Tracking, Head Pose Estimation, Gimbal Mechanism, Ministerial Decision No. 1275, TeleGimbal Startup.

Résumé

Ce mémoire présente la conception et l'implémentation du système *TeleGimbal*, un bras robotique rotatif intelligent basé sur l'intelligence artificielle pour le suivi intelligent du visage. Le système vise à résoudre le problème de perte de cible causé par le mouvement continu du conférencier dans les salles de classe et les amphithéâtres, où les caméras fixes ou les opérateurs humains ne parviennent pas à assurer un suivi fiable et économique.

Le système repose sur une webcam intégrée avec des algorithmes de vision par ordinateur OpenCV. La détection des visages est assurée par le modèle YuNet et la reconnaissance par le modèle SFace. À partir des coordonnées visuelles traitées, le système envoie des commandes de contrôle en temps réel via un microcontrôleur Arduino UNO à deux moteurs pas-à-pas NEMA 17 (axe Pan pour la rotation horizontale, axe Tilt pour l'inclinaison verticale) via des pilotes A4988 avec micro-pas 1/16, assurant un mouvement fluide, silencieux et sans vibrations.

Cinq défis techniques majeurs ont été documentés et résolus : la résonance acoustique (résolue par l'activation du micro-pas), le dépassement cinétique (résolu par une limite de vitesse stricte et une zone de tolérance de 50 pixels), les vibrations induites par l'intelligence artificielle (résolues par un filtre d'hystérésis), la désactivation d'urgence en cas de perte de cible (résolue par un arrêt d'urgence), et la gestion thermique (résolue par calibrage VREF à 0,8 V).

Les résultats expérimentaux démontrent une excellente efficacité opérationnelle avec un taux d'images entre 27,75 et 29,79 images par seconde, une latence inférieure à 50 millisecondes, une confiance de reconnaissance dépassant 0,7, et une capacité de mesure d'angle d'inclinaison jusqu'à 92,14. Sur le plan commercial, l'étude de faisabilité démontre une marge bénéficiaire brute de 71,6%, un seuil de rentabilité à 41 unités, et un retour sur investissement atteignant 133,6% pour 200 unités vendues. Ce prototype académique a été développé dans le cadre du Décret Ministériel 1275 (parcours Startup) pour devenir un produit commercial viable nommé *TeleGimbal*.

Mots-clés : Vision par ordinateur, Suivi de visage, Estimation de la pose de la tête, Mécanisme Gimbal, Décret Ministériel 1275, Startup TeleGimbal.

الملخص

TeleGimbal هو ذراع روبوتية دوارة ذكية تعمل بالذكاء الاصطناعي لتتبع الوجوه. يهدف النظام إلى حل مشكلة فقدان الهدف الناتجة عن الحركة المستمرة للمتحدثين داخل القاعات الدراسية والمحاضرات، حيث غالبًا ما تفشل الكاميرات الثابتة أو المشغلون البشريون في تتبع الحركة بشكل موثوق وفعال من حيث التكلفة.

يعتمد النظام على خوارزميات الرؤية الحاسوبية باستخدام مكتبة **OpenCV**، حيث يتم كشف الوجوه باستعمال نموذج **YuNet** والتعرف عليها باستخدام نموذج **SFace**. وبناءً على الإحداثيات البصرية الناتجة عن معالجة الصور، يقوم النظام بإرسال أوامر تحكم آنية عبر متحكم **Arduino UNO** لتشغيل محور الدوران الأفقي (**Pan**) ومحور الميل العمودي (**Tilt**) باستخدام محركات خطوية من نوع **NEMA 17**، يتم التحكم بها عبر مشغلات **A4988** مع تفعيل خاصية التقسيم الدقيق **1/16** (**Microstepping**)، مما يوفر حركة سلسة وصامتة وخالية من الاهتزازات.

تم توثيق ومعالجة خمسة تحديات هندسية رئيسية، وهي: ظاهرة الرنين الصوتي، التي تم حلها من خلال تفعيل التقسيم الدقيق؛ والتجاوز الحركي الناتج عن معالجة الذكاء الاصطناعي، والذي تم حله بتطبيق منطقة سماح قدرها 50 بكسل وتحديد السرعة القصوى؛ والاهتزازات الميكانيكية، التي تمت معالجتها عبر تعديلات فيزيائية وزيادة ارتفاع التثبيت؛ وفقدان الهدف، الذي تمت معالجته بإضافة نظام أمان يقوم بإرسال أمر توقف فوري؛ وارتفاع درجة حرارة المحرك الخطوي، الذي تم حله من خلال معايرة قيمة **VREF** إلى **0.8 فولت**.

أظهرت النتائج التجريبية كفاءة تشغيلية عالية، حيث حقق النظام معدل إطارات يتراوح بين **27.75** و**29.79 إطارًا في الثانية**، ومستوى ثقة في التعرف على الوجوه يتجاوز **0.7**، وزمن استجابة أقل من **50 ميلي ثانية**، ودقة في قياس الزوايا تصل إلى **92.14 درجة**. ومن الناحية التجارية، أظهرت دراسة الجدوى تحقيق هامش ربح إجمالي قدره **71.6%**، ونقطة تعادل عند بيع **41 وحدة**، وعائد على الاستثمار يصل إلى **133.6%** عند بيع **200 وحدة**.

تم تطوير هذا النموذج الأكاديمي في إطار **القرار الوزاري رقم 1275**، ويُقترح تحويله إلى منتج تجاري قابل للتسويق ضمن مشروع المؤسسة الناشئة **TeleGimbal**.

الكلمات المفتاحية: الرؤية الحاسوبية، تتبع الوجوه، تقدير وضعية الرأس، آلية الجيمبال، القرار الوزاري رقم 1275، المؤسسة الناشئة TeleGimbal.

List of Abbreviations and Acronyms

AI	Artificial Intelligence
CV	Computer Vision
ML	Machine Learning
DL	Deep Learning
CNN	Convolutional Neural Network
DNN	Deep Neural Network
GPU	Graphics Processing Unit
CPU	Central Processing Unit
FPS	Frames Per Second
RGB	Red Green Blue
ROI	Return on Investment
HOG	Histogram of Oriented Gradients
SIFT	Scale-Invariant Feature Transform
SURF	Speeded-Up Robust Features
YOLO	You Only Look Once
OpenCV	Open Source Computer Vision Library
GUI	Graphical User Interface
USB	Universal Serial Bus
ONNX	Open Neural Network Exchange
COM	Communication Port
IoT	Internet of Things
PC	Personal Computer
BOM	Bill of Materials
PTZ	Pan-Tilt-Zoom
DOF	Degrees of Freedom
PCB	Printed Circuit Board
PWM	Pulse Width Modulation
STEP	Step Pulse Signal
DIR	Direction Signal
VREF	Reference Voltage
SDK	Software Development Kit
API	Application Programming Interface
UART	Universal Asynchronous Receiver-Transmitter
A4988	Microstepping Motor Driver Module
NEMA	National Electrical Manufacturers Association
PID	Proportional-Integral-Derivative
JVM	Java Virtual Machine
IDE	Integrated Development Environment

Contents

Practical Part

3.5 Market Study 50
3.6 Microenvironment Analysis 51

List of Tables

List of Figures

General Introduction

The rapid evolution of artificial intelligence, computer vision, and robotics has enabled the development of autonomous systems capable of perceiving their environment and executing real-time actions. One of the most significant applications of these combined technologies is intelligent, dynamic subject tracking in unstructured environments.

In educational settings, such as classrooms and lecture halls, lecturers move naturally while presenting. In conference environments, speakers change dynamically and move across stages. Traditional tracking solutions rely on static cameras or human operators. Static cameras fail to follow moving speakers, resulting in poor video quality and the loss of important visual context. Human operators, on the other hand, introduce slow reaction times, fatigue, high operational costs, and a lack of scalability. The COVID-19 pandemic further accelerated the need for automated, reliable, and affordable tracking solutions for online education, hybrid conferences, and remote collaboration.

However, existing commercial products—such as those offered by DJI, OBSBOT, and Logitech—remain prohibitively expensive for many educational institutions and small-to-medium enterprises, particularly in developing countries. Moreover, traditional vision-only tracking frameworks frequently suffer from target loss caused by continuous speaker motion, partial occlusions, and dynamic illumination changes.

Problem Statement

This thesis addresses the following central research problem: How to design, implement, and validate an autonomous, low-cost, vision-based robotic tracking system capable of operating reliably in real time under dynamic environmental conditions without requiring expensive hardware or permanent human intervention?

To answer this central question, four specific sub-questions are formulated:

1. **Perception:** How can lightweight deep neural network models for visual face detection and recognition be effectively integrated into a real-time tracking pipeline that runs efficiently on standard processing units without requiring a dedicated graphics processing unit?

2. **Control and Actuation:** How can pixel-level tracking errors derived from computer vision

be translated into smooth, stable, and responsive mechanical motion using a microcontroller (Arduino UNO) and a low-cost pan-tilt stepper motor mechanism?

3. **Hardware and Engineering:** How can physical engineering challenges—such as acoustic resonance, kinetic overshoot, AI-induced jitter, and thermal management—be resolved to achieve professional-grade tracking performance?

4. **Commercialization and Viability:** How can the academic prototype be successfully transformed into a commercially viable product—specifically the *TeleGimbal* startup under the Ministerial Decree 1275 framework—featuring a sustainable business model, competitive pricing, and a positive return on investment?

These questions collectively define the problem statement of this thesis: bridging the gap between expensive, high-performance commercial tracking systems and affordable, robust vision-based solutions that can be deployed widely in educational and professional environments.

Main Objectives

The primary objective of this thesis is to engineer a functional, cost-effective vision-based robotic tracking prototype that overcomes the limitations of traditional static camera systems. The secondary objective is to validate the commercial feasibility of this prototype under the framework of Ministerial Decree 1275, transforming the academic research into a viable startup venture (*TeleGimbal*) with a validated business plan and go-to-market strategy.

Thesis Structure

To address the research questions and achieve the stated objectives, this thesis is organized into two main parts, comprising four chapters as follows.

Part I: Theoretical Foundations and Business Feasibility

This part establishes the theoretical and commercial groundwork for the TeleGimbal system. It is divided into two chapters.

- **Chapter 1: State of the Art in Vision and Robotics** establishes the theoretical foundation of the project. It explores the fundamental concepts of robotic manipulators, computer vision workflows (specifically face detection and recognition), and the integration of vision-based tracking architectures.
- **Chapter 2: The Startup Project and Business Plan** transitions the technical research into a commercial framework. It presents the TeleGimbal startup project, detailing the market study, competitive analysis, business model, and financial viability required to deploy the system in educational and professional environments.

Part II: Practical Implementation, Validation, and Future Perspectives

This part presents the complete practical implementation of the TeleGimbal system, its experimental validation, and the general conclusion with future work. It is divided into two chapters.

- **Chapter 3: System Design, Hardware Architecture, Software Implementation, and Experimental Validation** details the selection, assembly, and integration of the electronic components, including the Arduino UNO, camera module, NEMA 17 stepper motors, A4988 drivers, and the custom 3D-printed pan-tilt mechanism. This chapter also presents the coding architecture, the vision-based tracking algorithms (YuNet for face detection and SFace for face recognition), the Bang-Bang velocity controller with a tolerance box, and the performance evaluations conducted to measure tracking accuracy, frame rate, latency, and system robustness in real-world academic environments.
- **Chapter 4: General Conclusion and Future Perspectives** summarizes the main findings of the research, discusses the technical and commercial limitations encountered during the development of the TeleGimbal system, and proposes future perspectives for both technological improvement and market expansion. These future perspectives include audio-based sound source localization for bimodal tracking, wireless communication using ESP32, standalone operation without a host computer using edge AI devices such as Raspberry Pi, solar power capability for off-grid deployment, multi-target tracking and automatic speaker switching, cloud integration for remote management, and assistive healthcare applications.

THEORETICAL PART

Literature Review, Background, and Concepts

Chapter 1

State of the Art in Vision, and Robotics

1.1 Introduction

Over the last decade, artificial intelligence has experienced a significant transformation, evolving from theoretical concepts into practical technologies that support a variety of autonomous systems. This progression has been driven by the convergence of related scientific fields, notably Computer Vision and Robotics, which together enable machines to perceive their surroundings, interpret complex sensory information, and perform context-aware actions. This integration marks a fundamental shift from passive data processing to active, intelligent engagement with real-world environments [1], [2], [3]. Central to this advancement is Computer Vision, which endows machines with the ability to derive meaningful information from visual data such as images and video streams. Leveraging sophisticated algorithms, systems can accomplish tasks including object detection, facial analysis, and motion tracking with growing precision and reliability. Concurrently, Robotics provides essential mechanical and control frameworks that transform these perceptual insights into tangible actions. The interplay between perception and actuation underpins intelligent systems capable of real-time responsiveness and adaptive behavior in dynamic settings [1], [2], [3], [4]. Despite these technological gains, many traditional monitoring systems still depend on manual operation. In settings like classrooms, lecture halls, and conferences, camera control is frequently assigned to human operators who are responsible for tracking subjects and maintaining visual focus. This manual approach introduces several drawbacks, such as limited reaction times, vulnerability to fatigue, and increased operational expenses. Moreover, it lacks scalability and consistency, especially in scenarios requiring prolonged or multi-subject monitoring [2], [4].

To address these challenges, intelligent tracking systems have been developed as an advanced alternative, integrating multiple sensing and actuation modalities. These systems combine Computer Vision for visual perception, audio signal processing for locating sound sources, and robotic platforms for dynamic camera control. Through this multimodal integration, the system can detect and identify human subjects, recognize active speakers, and continuously adjust its orientation to optimize tracking performance. This capability substantially improves system autonomy and reliability in complex environments [1], [3], [4], [5].

A crucial factor for the success of these systems is their capacity to function in real time while maintaining robustness amid environmental variability. This requirement demands the creation of efficient perception algorithms alongside sophisticated control strategies that ensure smooth and accurate robotic movements. Additionally, the close coupling between sensing and actuation presents challenges concerning synchronization and latency, making feedback-driven control mechanisms indispensable for stable and responsive system operation [2], [4], [5]. This chapter lays a comprehensive groundwork for understanding the design and implementation of intelligent tracking systems. It covers fundamental topics including robotic arm kinematics and configurations, Computer Vision processing workflows, face detection and recognition techniques, audio-based localization methods, and integrated tracking frameworks. Collectively, these elements provide the theoretical and practical foundation for the advanced system architecture discussed in later chapters [1] [2] [4] | Hartley & Zisserman, 2004 | [3].

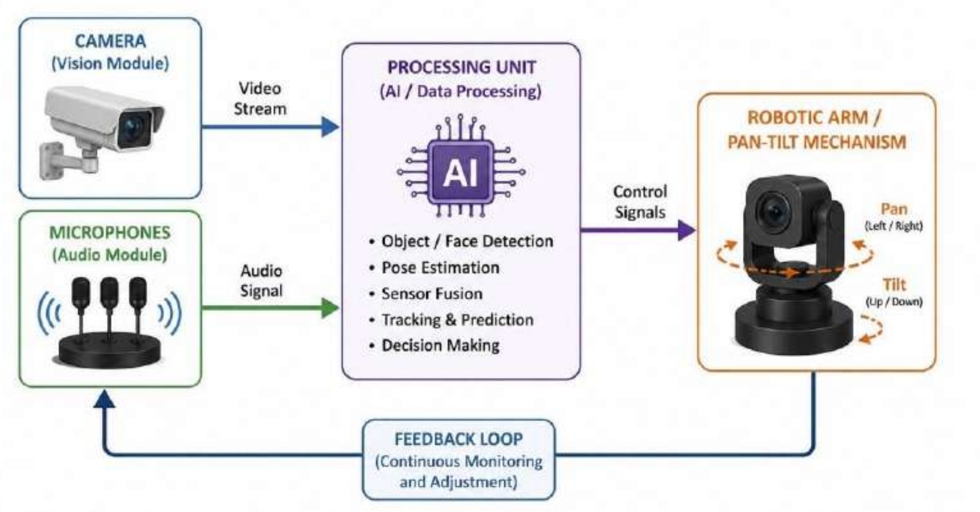


Figure 1.1: Typical Intelligent Tracking System Architecture.

1.2 Robotic Systems

1.2.1 Definition of a Robotic Arm

A robotic arm is an advanced electromechanical system engineered to replicate the kinematic behavior and functional capabilities of the human arm. It is composed of a sequence of rigid bodies, known as links, interconnected through mechanical joints that enable controlled motion. These joints may be classified as revolute joints, which provide rotational movement, or prismatic joints, which allow linear displacement. Through the coordinated interaction of these components, robotic arms are capable of performing a wide range of tasks within a predefined workspace with high levels of accuracy, repeatability, and reliability. Figure 2 presents a typical example of a robotic arm architecture commonly used to illustrate the fundamental elements of robotic manipulators. It should be noted, however, that this diagram is intended solely as a representative example and does not constitute a universal or standardized architecture. In practice, robotic systems may adopt diverse mechanical configurations and control structures depending on their intended applications, performance requirements, and design constraints [6], [7].

From a kinematic standpoint, a robotic arm is primarily characterized by its Degrees of Freedom (DOF), which represent the number of independent parameters required to define the position and orientation of the manipulator. The number of degrees of freedom directly influences the robot's mobility, dexterity, and ability to execute complex trajectories. Robotic arms with a higher number of DOF can perform more sophisticated movements and adapt more effectively to dynamic and constrained environments. Consequently, the kinematic design of the manipulator plays a fundamental role in determining its operational capabilities and the range of tasks it can accomplish [7], [8].

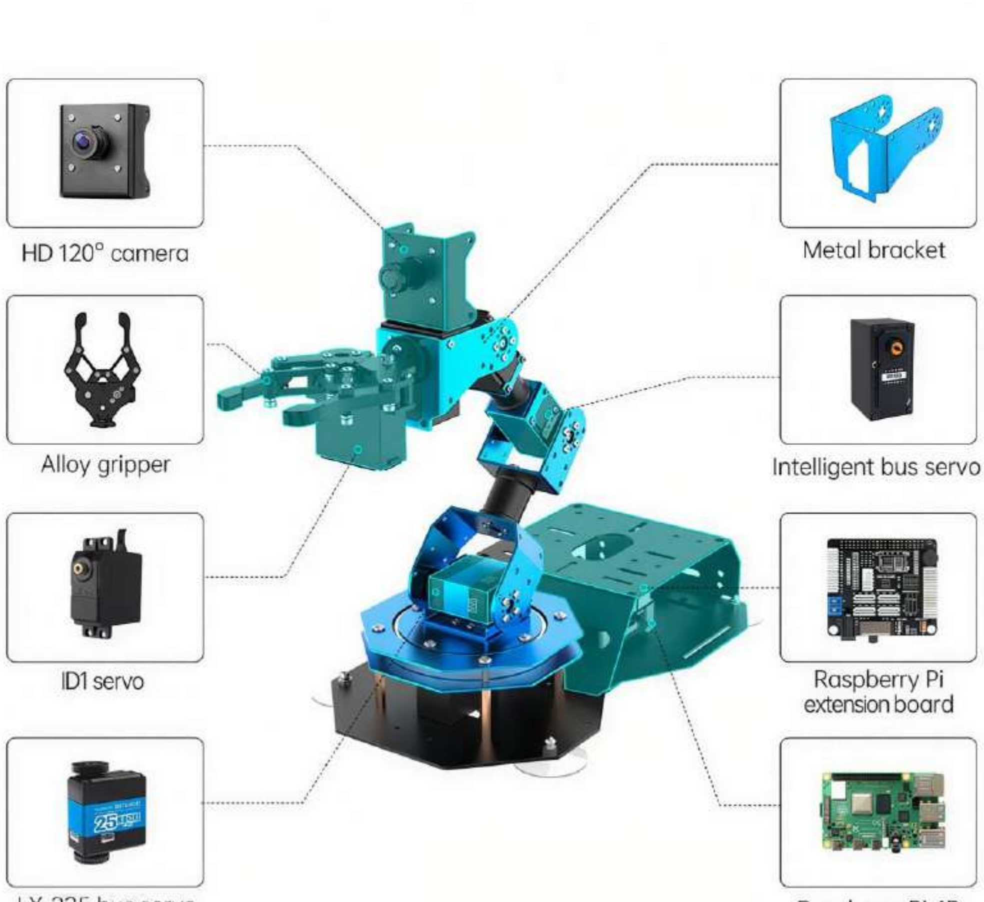


Figure 1.2: Illustrative Example of a Robotic Arm Architecture and Its Main Components [9]

Joint actuation is achieved through various mechanisms, a vital aspect of the robotic arm's performance. These actuators encompass electric motors—including servo and stepper motors—as well as hydraulic and pneumatic systems, each offering distinct trade-offs in power output, precision, and response dynamics. Contemporary robotic applications predominantly employ servo motors due to their superior accuracy, fine-grained controllability, and seamless integration with real-time control frameworks. Embedded control units govern actuator operation, ensuring precise joint positioning, smooth trajectory execution, and consistent repeatability under diverse operational conditions [2], [4].

Beyond mechanical and actuation components, modern robotic arms are increasingly integrated with an array of sensors and advanced control algorithms facilitating closed-loop, feedback-driven operation. Sensors continuously monitor critical system variables such as joint angles, angular velocities, and applied forces, enabling the control system to dynamically modulate behavior in response to both internal disturbances and environmental fluctuations. This feedback control paradigm substantially enhances the system's robustness, accuracy, and stability, particularly in scenarios characterized by uncertainty and variability [3], [4].

Within the scope of Computer Vision and Intelligent Tracking Systems, the robotic arm functions as the pivotal physical interface connecting perception to action. Vision-based algorithms detect, localize, and track targets—such as human faces or moving objects—in the environment. The robotic arm translates these perceptual inputs into precise kinematic commands, thereby maintaining continuous alignment with the target. This seamless integration of sensing and actuation is essential for achieving real-time autonomous tracking capabilities [1], [3], [4].

In practical deployments, the robotic arm typically manipulates the orientation and spatial positioning of a camera or sensor platform. By continuously adjusting its pose, the system ensures persistent target presence within the sensor's field of view, even amid target motion or environmental variability. This functionality imposes stringent requirements on system performance, including smooth, uninterrupted motions, minimal latency, and high positional accuracy, underscoring the necessity for precise actuation and efficient real-time controlmethodologies [2], [4]. From a holistic standpoint, the robotic arm should be regarded not merely as a mechanical manipulator but as an integral element of an intelligent autonomous system. It epitomizes the convergence of perception, decision-making, and physical interaction, empowering machines to engage with their environment purposefully and adaptively. This comprehensive perspective highlights the critical role of robotic arms in advancing next-generation intelligent robotic systems [2], [3], [4].

Component	**Function**	**Example**
Sensors	Data acquisition	Camera, Microphone
Actuators	Motion execution	Servo motors
Controller	Decision making	Microcontroller

Table 1.1: Fundamental Components of Robotic Systems

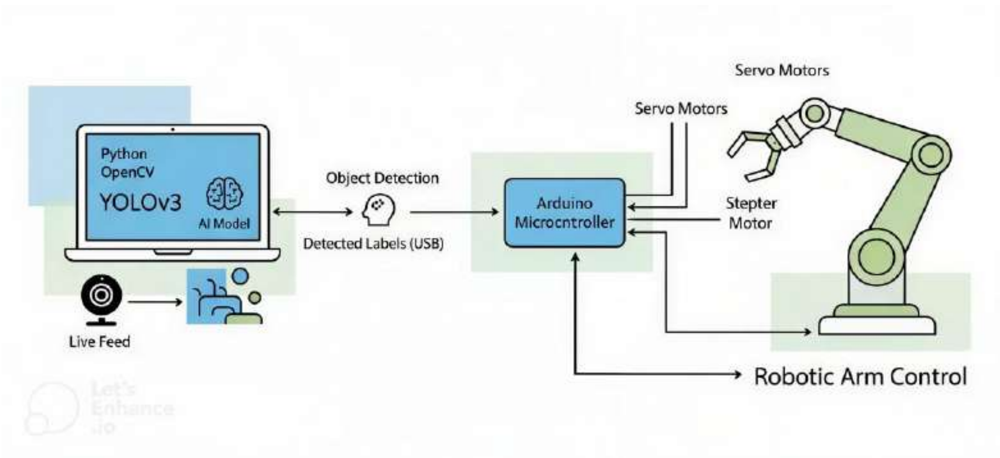


Figure 1.3: Vision-Based Robotic Arm Tracking System [10]

1.2.2 Classification of Robotic Arms and Their Relevance to Computer Vision and Intelligent Tracking Systems

The classification of robotic arms is foundational for selecting and optimizing manipulator architectures tailored to specific application domains, particularly those involving Computer Vision and intelligent tracking. Each class of robotic arm exhibits distinct kinematic configurations, operational capabilities, and mechanical constraints that influence system performance in perception-driven, real-time autonomous environments [1], [2].

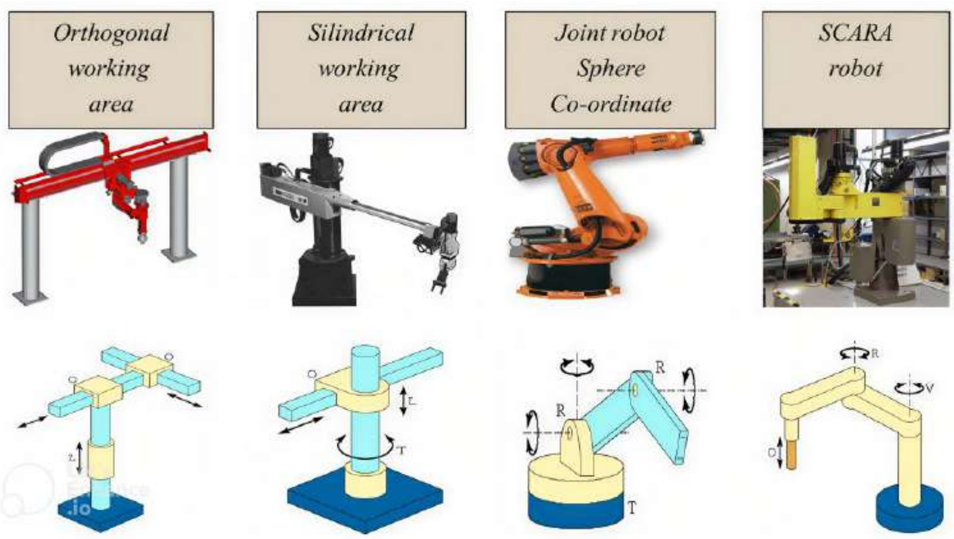


Figure 1.4: Basic Types of Industrial Robots [11]

1.2.3 Articulated Robotic Arms

Articulated robotic arms constitute one of the most prevalent and versatile categories of manipulators, distinguished by their series of interconnected rotary joints, typically encompassing three to six degrees of freedom (DOF). This kinematic architecture closely mimics the human arm, facilitating smooth, continuous, and highly dexterous motion within a three-dimensional operational workspace. The anthropomorphic design inherently supports complex spatial trajectories and nuanced orientation control, which are indispensable for tasks demanding high precision and flexibility.

In intelligent tracking systems that integrate Computer Vision, articulated arms offer significant advantages due to their rotational mobility and extensive maneuverability. Their capability to execute real-time, multidimensional adjustments allows continuous alignment of vision sensors with dynamically moving targets, a critical requirement in applications such as face tracking, gesture recognition, and interactive robotics [1], [2], [4].

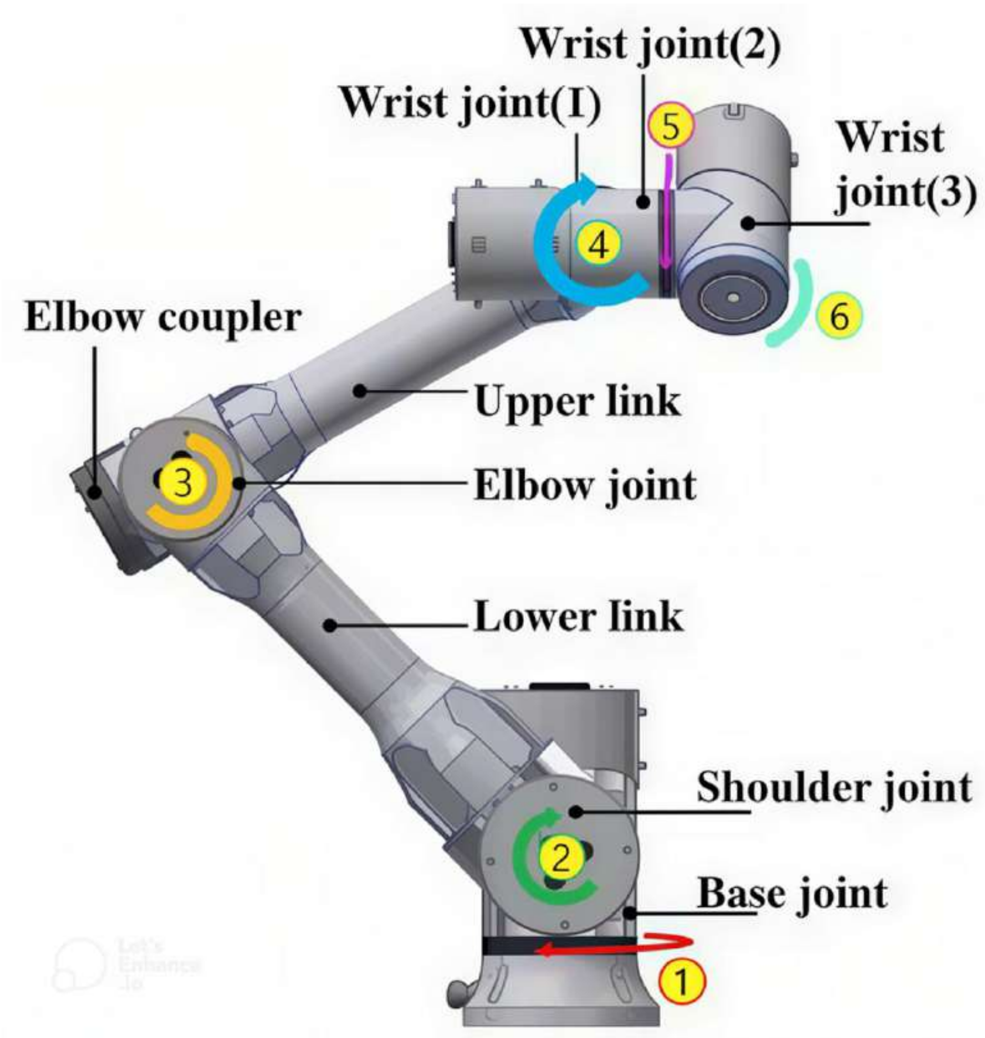


Figure 1.5: Structure of an Articulated Robotic Arm [12]

SCARA Robots

Selective Compliance Assembly Robot Arm (SCARA) robots are characterized by a specialized kinematic structure that provides selective compliance within the horizontal plane while maintaining rigidity along the vertical axis. This configuration enables rapid, high-precision planar movements, making SCARA robots particularly well-suited for repetitive assembly tasks.

Despite their speed and accuracy, SCARA robots exhibit inherent limitations in Computer Vision-based tracking applications due to their restricted motion, which is predominantly confined to two-dimensional planes. This constraint hampers their ability to perform continuous spatial orientation adjustments necessary for effective tracking of three-dimensional, dynamically moving targets [1], [2], [5].

Figure 1.6: Robot Arm (SCARA) [13]

Cartesian Robotic Arms

Cartesian robotic arms, also known as gantry or linear robots, operate through translational motion along three mutually orthogonal linear axes—X, Y, and Z. Their straightforward kinematic configuration simplifies mathematical modeling and control strategies, facilitating reliable execution of linear, repetitive tasks within structured environments

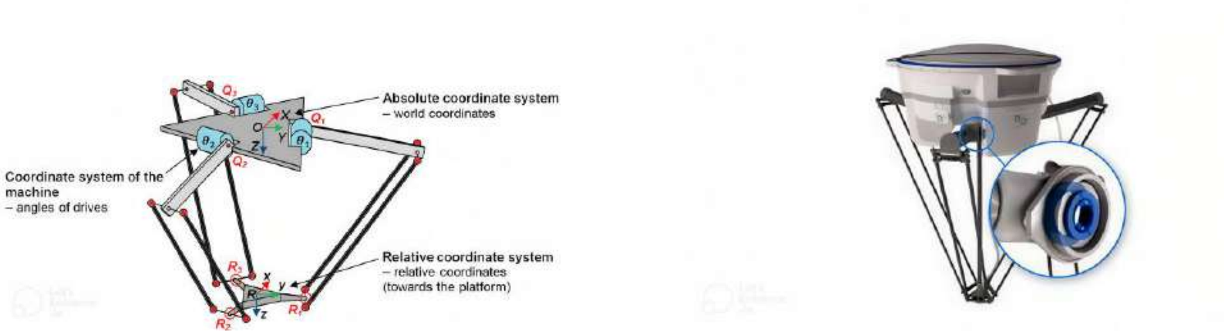
In Computer Vision applications, Cartesian robots are typically deployed in controlled settings where object positions are predefined. Their lack of rotational degrees of freedom limits their adaptability in dynamic tracking systems requiring continuous reorientation [1], [2].

Figure 1.7: Cartesian Robotic Arms [14]

Delta Robots

Delta robots feature a parallel kinematic structure composed of multiple arms connected to a fixed base, forming a closed-loop mechanism. This design enables extremely rapid and precise movements, especially for lightweight objects, due to reduced moving mass and high dynamic efficiency.

However, their workspace is limited, and motion is restricted to predefined volumes, which reduces their adaptability for vision-based tracking in dynamic environments requiring wide spatial coverage [1], [3].



(a) Coordinate system of Delta robot

FST Robotics

(b) Absolute and relative coordinate systems

Acrome Robotics

Figure 1.8: Delta robot coordinate systems

1.2.4 Collaborative Robots (Cobots)

Collaborative robots (cobots) are designed to operate safely alongside humans in shared environments. They integrate advanced sensors and intelligent control systems that enable adaptive and responsive behavior. When combined with computer vision techniques, cobots can detect human presence, interpret motion, and dynamically adjust their actions in real time. This makes them particularly suitable for interactive tracking and a wide range of human-centered robotics applications where safety and close human–machine collaboration are required [4], [15], [16].

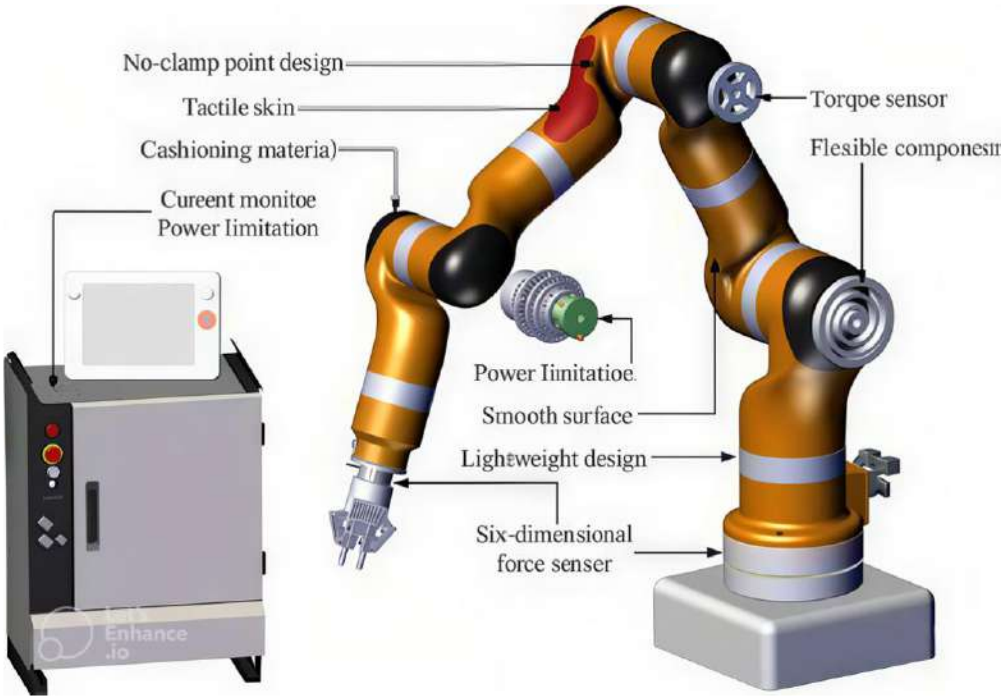


Figure 1.9: Collaborative Robots (Cobots) [17]

Gantry Robots

Gantry robots are large-scale robotic systems operating on fixed frames using linear axes to cover extensive workspaces. They are designed for heavy payloads and large-area industrial tasks. However, their purely translational motion limits their ability to perform real-time orientation adjustments required in vision-based tracking systems [1]. The integration of robotic manipulators within Computer Vision–based intelligent tracking systems necessitates a deliberate and technically informed selection of the robotic arm configuration, as this choice fundamentally governs the system's overall efficacy and responsiveness. Such selection must be predicated on a comprehensive assessment of multiple design criteria, encompassing degrees of freedom (DOF), kinematic architecture, motion accuracy, response latency, and adaptability to dynamic or unstructured environments [2], [4].

From both kinematic and operational standpoints, the number of degrees of freedom represents a paramount design parameter. A greater DOF enhances the manipulator's spatial flexibility, allowing the end-effector to access a broader range of positions and orientations. This attribute is particularly critical in vision-guided tracking contexts, where targets exhibit unpredictable and complex motion trajectories. Articulated robotic arms, characterized by multiple rotational joints, are thus optimally suited for such applications [4], [18].

Beyond mechanical configuration, dynamic performance considerations are equally vital for intelligent tracking applications. Real-time vision systems demand rapid sensory data processing coupled with immediate actuation to preserve tracking fidelity and system stability. Latencies between perception and motor execution can result in tracking inaccuracies or loss of target continuity [5]. Another critical factor in evaluating robotic manipulators for Computer Vision integration is the degree of interoperability between the mechanical platform and the computational intelligence framework. Modern robotic systems increasingly employ closed-loop control architectures wherein visual feedback is continuously assimilated to iteratively refine motion commands [2], [5].

In summary, comparative analysis reveals that no single robotic arm configuration universally satisfies all performance criteria; instead, selection must be aligned with specific application demands and system constraints. However, for Computer Vision–driven intelligent tracking, manipulators exhibiting high degrees of freedom, versatile kinematic adaptability, and advanced real-time control—exemplified by articulated robotic arms—are generally favored [4], [18].

Type	**Degrees of Freedom (DOF)**	**Advantages**	**Applications**
Articulated	3-6 DOF	High flexibility, 3D motion	Visual tracking, Industry
SCARA	4 DOF	Fast, high precision in horizontal plane	Assembly, Electronics
Cartesian	3 DOF	Simple, linear motion	3D printing, Pick and place
Delta	3-4 DOF	Very fast, lightweight	Packaging, Pharmaceuticals
Collaborative (Cobot)	6-7 DOF	Human-safe, adaptive	Healthcare, HRI
Gantry	3-5 DOF	Heavy payload, large workspace	Logistics, Heavy manufacturing

Table 1.2: Classification of Robotic Arms and Their Characteristics

1.3 Computer Vision

1.3.1 Definition of Computer Vision

Computer Vision is an advanced subfield of Artificial Intelligence dedicated to enabling computational systems to autonomously interpret visual data obtained from images and video streams. Its core aim is to replicate essential human visual cognition, allowing machines to perform nuanced analysis of visual inputs and derive semantically rich, context-aware interpretations across diverse visual applications, including but not limited to object detection, motion tracking, image segmentation, and holistic scene understanding **zhao2023**.

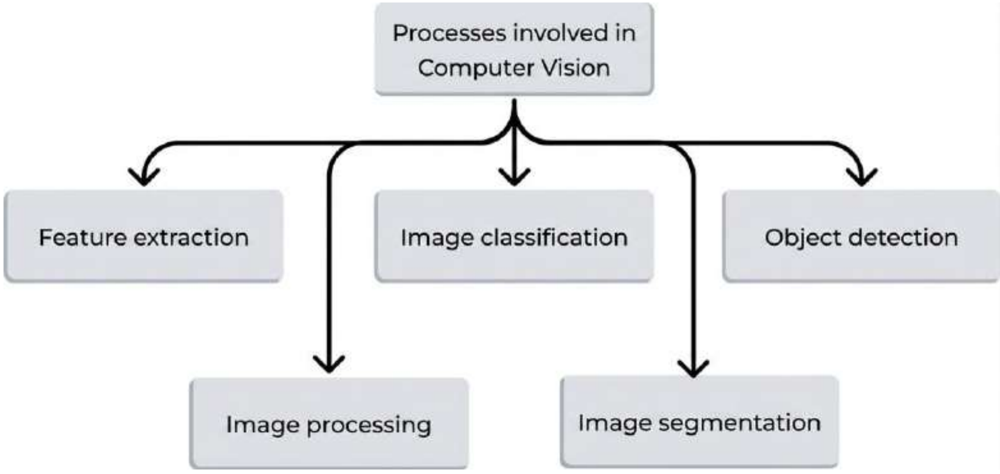


Figure 1.10: Fundamental Processes in Computer Vision [19]

1.3.2 Computer Vision Pipeline

A typical Computer Vision system is architected as a hierarchical, multi-stage processing pipeline that systematically converts raw visual data into semantically enriched, high-level information suitable for analysis and decision-making. The pipeline initiates with the image acquisition phase, wherein visual inputs are captured from sensing devices [1], [20].

Subsequent to acquisition, the data undergo a series of preprocessing operations designed to enhance image quality and robustness. These operations typically include noise reduction techniques, illumination normalization, contrast enhancement, and geometric corrections [20], [21].

Following preprocessing, the pipeline advances to feature extraction, a pivotal stage wherein salient visual attributes—such as edges, corners, textures, gradients, or keypoints—are detected and encoded into structured representations [1], [21].

The extracted features are subsequently processed by higher-level modules dedicated to object detection, recognition, and classification. Leveraging learned models, statistical inference, or rule-based criteria, these modules identify and categorize objects or regions of interest [1], [22]. Collectively, this structured pipeline forms the foundational backbone of most Computer Vision frameworks, enabling automated image interpretation, real-time visual analysis, and intelligent decision-making across a diverse array of applications [1], [20], [22].

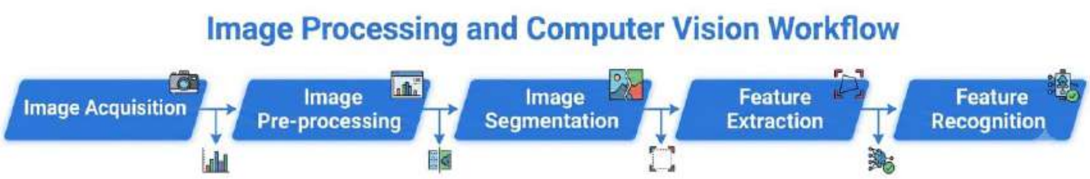


Figure 1.11: Image Processing and Computer Vision Workflow [23]

1.3.3 Challenges in Computer Vision

Despite significant advancements in the development of Computer Vision systems, numerous intrinsic challenges persist that constrain their robustness and efficacy in real-world deployments. The foremost challenge arises from variability in illumination conditions. Fluctuations in lighting intensity, the presence of shadows, and variations in exposure can substantially alter the visual characteristics of objects [1], [20].

Furthermore, complex and cluttered backgrounds introduce substantial visual noise and extraneous information, complicating the separation of relevant foreground objects from irrelevant scene elements. This issue is particularly pronounced in unconstrained environments [1], [21].

Real-time processing requirements impose additional constraints, especially in embedded or resource-constrained platforms. These constraints necessitate a careful balance between computational efficiency and algorithmic accuracy [22], [24].

Another significant challenge concerns the variability of object appearance due to changes in scale, viewpoint, orientation, partial occlusions, and diverse environmental conditions [1], [20], [24].

1.4 Face Detection and Face Recognition

1.4.1 Definition of face detection

Face detection constitutes a pivotal and foundational task within the broader domain of Computer Vision, entailing the automated identification and precise localization of human faces in digital images or video streams. At its core, this task involves determining the presence or absence of one or multiple faces within a given visual input and delineating their spatial extents. These spatial localizations are commonly represented using bounding boxes; however, more sophisticated methodologies extend beyond simple localization by identifying detailed facial landmarks such as eyes, nose, mouth, and other key points, facilitating richer geometric and semantic understanding of facial structures **zha23**. This capability is critical for enabling downstream facial analysis tasks that require fine-grained spatial information.

1.4.2 Applications and Importance

As a crucial preliminary stage in contemporary intelligent vision systems, face detection underpins a diverse array of higher-level applications, thereby serving as an indispensable enabler of advanced functionalities. These applications span a broad spectrum including, but not limited to, face recognition systems which authenticate identity, facial expression analysis for affective computing, gaze estimation relevant in human-computer interaction, biometric authentication frameworks ensuring security, surveillance infrastructures for public safety, human–robot interaction paradigms, and immersive augmented reality environments **jai24**. The reliability and accuracy of the face detection module directly influence the efficacy of these subsequent tasks; inaccuracies or failures at this stage propagate through the pipeline, substantially compromising the overall system performance and robustness. Hence, the development of robust face detection algorithms remains a critical research focus.

1.4.3 Historical Evolution of Face Detection Methods

Historically, the evolution of face detection methodologies has witnessed significant transformations aligned with broader advances in machine learning and computer vision. Early detection approaches predominantly utilized handcrafted feature extraction techniques combined with traditional classifiers. Notable examples include the Viola-Jones framework, which employed Haar-like features organized in a cascade classifier architecture, providing real-time performance under controlled conditions **vio23**. Similarly, the use of Histogram of Oriented Gradients (HOG) descriptors paired with classifiers such as Support Vector Machines (SVM) constituted a classical paradigm yielding considerable success in mid-2000s applications. Despite their efficiency and effectiveness in well-constrained environments, these classical methods exhibited limitations in generalizing to unconstrained, real-world scenarios characterized by diverse lighting, occlusions, and pose variations **dal23**.

1.4.4 Deep Learning and Modern Architectures

The advent and recent proliferation of deep learning architectures, particularly spanning the years 2023 to 2026, have precipitated a paradigm shift in face detection research and applications. Contemporary approaches predominantly leverage Convolutional Neural Networks (CNNs), which autonomously learn hierarchical, multi-scale, and highly discriminative feature representations directly from extensive annotated datasets, circumventing the need for manual feature engineering**he23**. More recently, transformer-based architectures—originally conceived for natural language processing tasks—have been adapted for computer vision, including face detection, offering enhanced capabilities in capturing global contextual dependencies and improving robustness under challenging conditions **dos24**. These state-of-the-art models demonstrate significant improvements in detection accuracy and resilience against common obstacles such as illumination changes, extreme head poses, partial occlusions, motion blur, low-resolution inputs, and cluttered backgrounds. The integration of multi-task learning frameworks that jointly optimize detection with landmark localization or attribute recognition further advances the robustness and versatility of face detectors **dos24**, **wan25** .

1.4.5 Challenges and Ongoing Research

Despite these notable advancements, face detection persists as a complex and unsolved challenge within artificial intelligence and computer vision domains. The intrinsic variability of human faces—encompassing factors such as dynamic facial expressions, aging effects, ethnic and cultural diversity, and accessories like glasses or masks—imposes substantial difficulty on detector generalization **li23**. Moreover, environmental conditions such as uneven lighting, shadows, and motion artifacts continue to degrade detection reliability. To address these issues, ongoing research efforts focus on enhancing real-time processing capabilities and computational efficiency to accommodate deployment on embedded systems and robotic platforms with limited hardware resources. Techniques such as model pruning, quantization, and knowledge distillation are actively investigated to balance the trade-off between accuracy and resource consumption **han24**. Additionally, the incorporation of synthetic data augmentation, domain adaptation strategies, and self-supervised learning paradigms are emerging trends aimed at improving model robustness and adaptability to diverse operational scenarios **chen26**.

1.4.6 Section Summary

In summation, face detection remains an indispensable and foundational step in any facial analysis pipeline, forming the bedrock upon which subsequent facial understanding tasks depend. The continuous evolution of face detection methodologies reflects the dynamic interplay between theoretical advances in artificial intelligence, algorithmic innovations in deep learning, and practical demands from real-world applications. As such, it remains one of the most intensely researched and impactful areas within computer vision and AI, with ongoing developments promising to further enhance the accuracy, efficiency, and applicability of face detection systems across diverse domains.

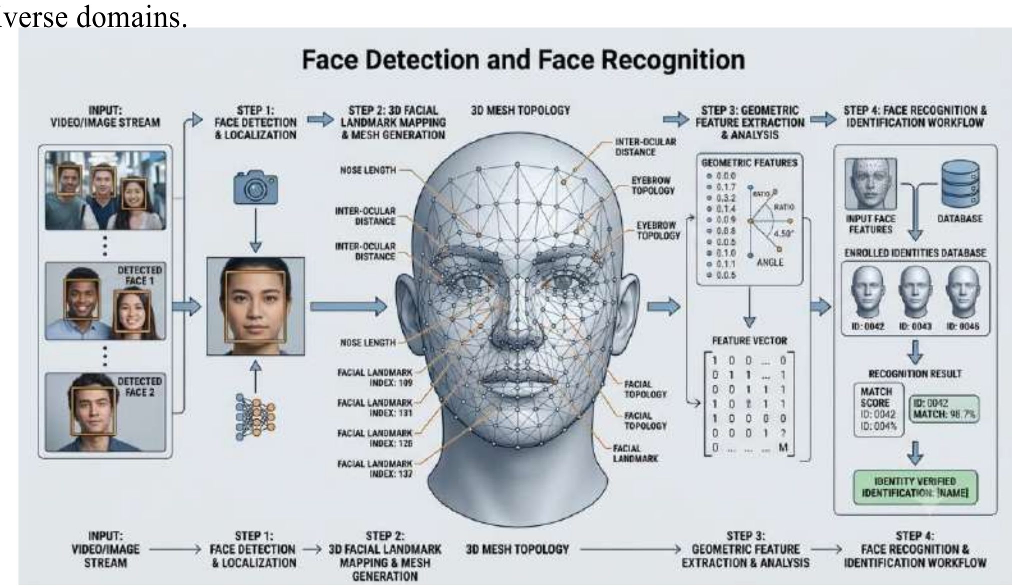


Figure 1.12: Workflow of face detection and face recognition based on facial landmark extraction and geometric feature analysis [25]

This diagram illustrates the stages of face recognition, which begin with acquiring images or video, followed by detecting and tracking the face's location, size, and pose. Next, face alignment standardizes the face position, and important features are extracted as feature vectors. Finally, these features are matched against a database of enrolled users to accurately identify the face **zhao23**.

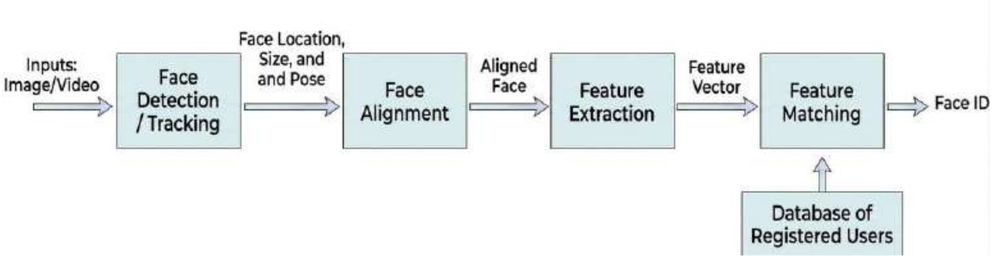


Figure 1.13: Face Recognition Processing Flow Diagram [26]

1.4.7 Importance of Face Detection

The significance of face detection extends across numerous scientific, industrial, and security-oriented applications. In intelligent surveillance systems, face detection enables automated monitoring, suspect identification, and anomaly tracking. Within biometric authentication frameworks, it acts as the gateway for facial recognition systems used in smartphones, banking systems, border control, and access management platforms. In robotics and autonomous systems, face detection facilitates natural human–robot interaction by allowing robots to identify and track users in real time. Furthermore, modern educational technologies, conferencing platforms, and social media applications rely heavily on face detection for user engagement, augmented reality filters, virtual attendance systems, and automatic camera framing.

From an Artificial Intelligence perspective, face detection constitutes a critical perception mechanism that bridges visual sensing and cognitive interpretation. Its effectiveness directly influences the performance of downstream modules such as face recognition, speaker tracking, behavioral analysis, and emotional understanding. Therefore, inaccuracies in face localization can propagate throughout the entire intelligent system, leading to degraded recognition accuracy and unstable tracking performance. For this reason, robust and real-time face detection algorithms are essential in modern AI-driven environments [27] [28].

1.4.8 Face Detection Techniques

Over the past decades, researchers have proposed a wide spectrum of methodologies for face detection, ranging from rule-based systems to sophisticated deep learning architectures. These techniques can generally be categorized into traditional handcrafted-feature approaches and modern data-driven learning paradigms.

Traditional methods primarily rely on manually designed features and statistical classification models. These approaches emphasize computational efficiency and were widely adopted before the emergence of deep learning. In contrast, modern approaches utilize neural networks capable of automatically learning hierarchical facial representations directly from training data. Although deep learning methods significantly improve robustness and accuracy, they generally require higher computational resources and large-scale annotated datasets [29] [30].

1.4.9 Classification of Face Detection Methods

Face detection methodologies are commonly classified into four major categories:

- Knowledge-Based Methods
- Feature Invariant Approaches
- Template Matching Methods
- Appearance-Based Methods

Each category addresses the face detection problem from a distinct theoretical and computational perspective.

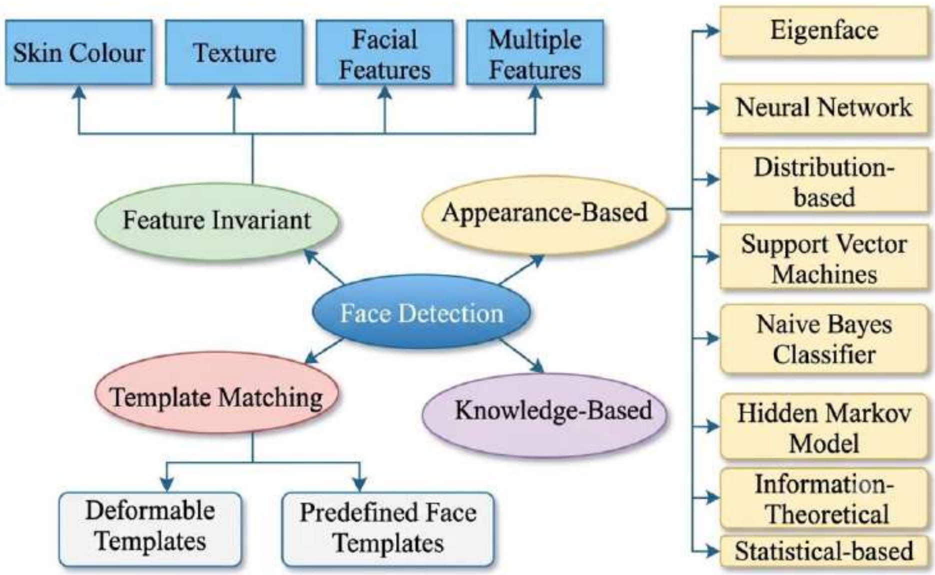


Figure 1.14: Categories of Face Detection Methods [31]

Knowledge-Based Methods

Knowledge-based approaches, also known as rule-based methods, utilize predefined human knowledge regarding facial structure and geometry. These techniques encode explicit rules describing the relationships between facial components such as the eyes, nose, mouth, and eyebrows. For example, a typical rule may specify that the eyes appear symmetrically above the nose, while the mouth lies beneath the nose region.

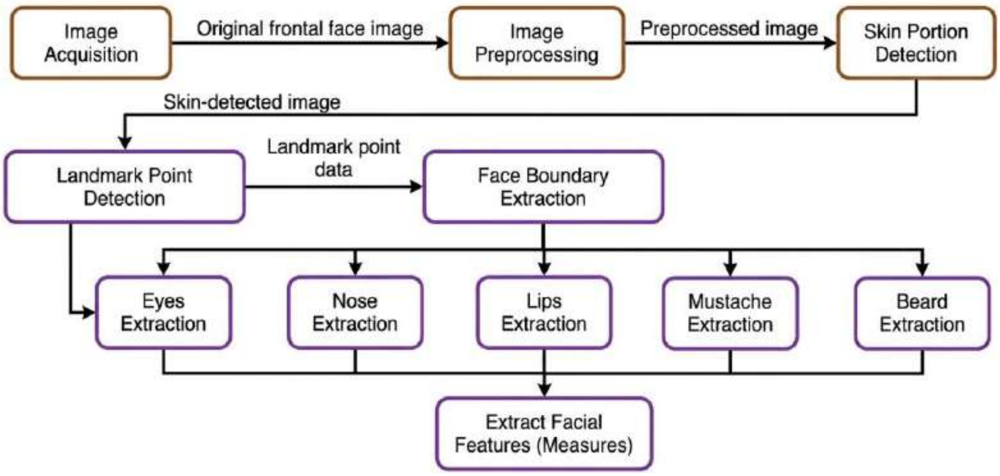


Figure 1.15: Knowledge-Based Face Detection and Landmark Extraction Flow [32]

The principal advantage of knowledge-based methods lies in their conceptual simplicity and low computational requirements. They are particularly effective in constrained environments where frontal facial poses and controlled illumination conditions are guaranteed. However, their performance deteriorates significantly when confronted with pose variations, facial occlusions, or complex backgrounds. Additionally, designing universal rules capable of handling the diversity of human faces remains an extremely difficult task. Consequently, purely rule-based systems are rarely employed in modern intelligent systems | Yang et al., 2002 |.

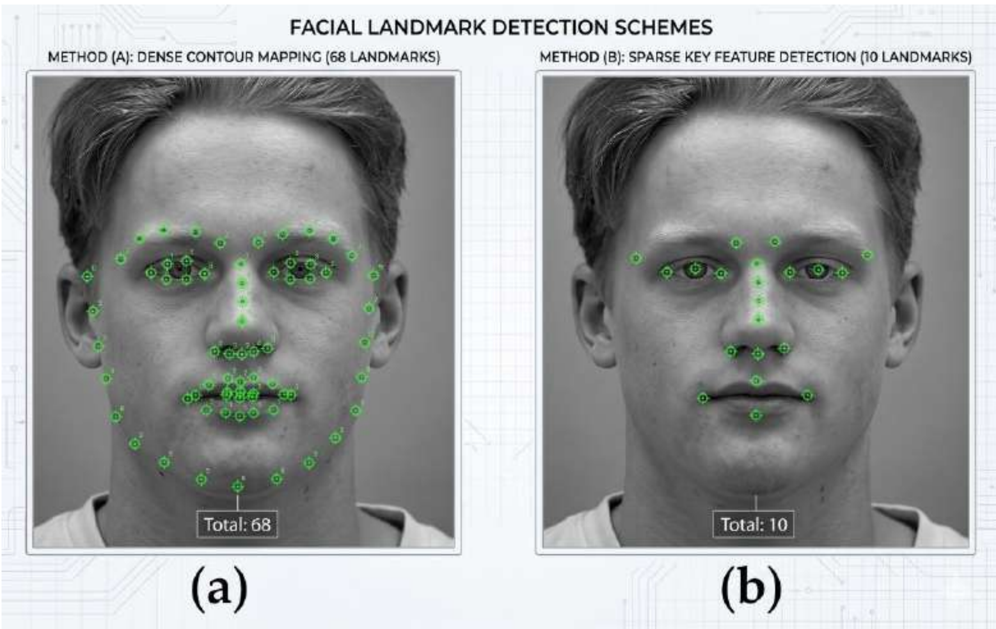


Figure 1.16: Dense and Sparse Facial Landmark Detection Models [33]

Feature Invariant Approaches

Feature invariant methods attempt to identify structural facial characteristics that remain relatively stable under varying imaging conditions. Instead of relying on holistic face representations, these approaches focus on invariant features such as skin color, facial texture, edges, contours, or geometric relationships.

Skin-color segmentation techniques, for instance, exploit chromatic information in color spaces such as RGB, HSV, or YCbCr to isolate probable facial regions. Edge-based methods analyze gradients and contour distributions to approximate facial structures. The primary advantage of feature invariant approaches is their improved robustness against partial occlusions and moderate pose variations.

Nevertheless, these techniques are highly sensitive to illumination changes and environmental conditions. Skin-color models may fail under non-uniform lighting or across different ethnic skin tones. Similarly, contour-based methods often generate false detections in cluttered scenes containing face-like structures. Therefore, feature invariant techniques are typically combined with additional classification stages to enhance reliability [28] [27].

Template Matching Methods

Template matching methods rely on predefined facial templates used to compare and locate face-like regions within an image. A template may represent the entire face or specific facial components such as the eyes or mouth. During detection, correlation operations are performed between the template and different image regions to identify the highest similarity scores.

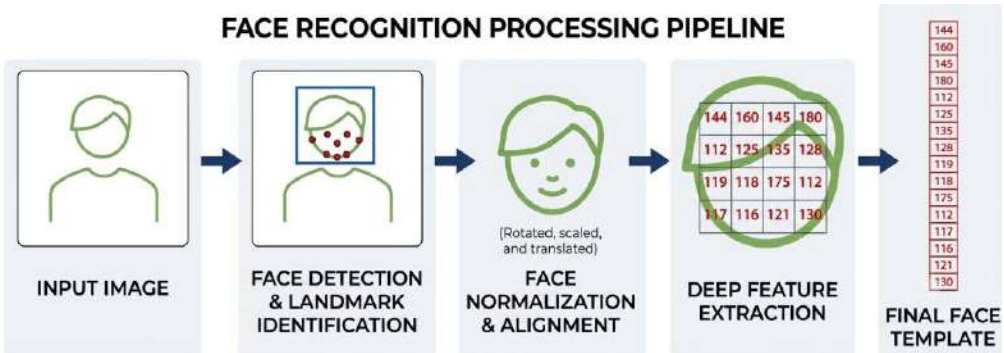


Figure 1.17: Template Matching Face Detection Method [34]

These methods can be categorized into fixed-template and deformable-template approaches. Fixed-template techniques utilize rigid predefined patterns and are computationally simple. In contrast, deformable templates allow adaptive transformations to accommodate moderate variations in facial shape and expression.

Although template matching provides intuitive and interpretable detection mechanisms, its performance is often constrained by sensitivity to scale, rotation, and illumination variations. Furthermore, exhaustive template comparisons increase computational cost, particularly in high-resolution video streams. As a result, template-based approaches have largely been superseded by machine learning and deep learning techniques in contemporary Computer Vision systems [35].

Appearance-Based Methods

Appearance-based methods represent a major evolution in face detection research. Instead of manually defining facial characteristics, these techniques learn discriminative facial representations directly from training datasets using statistical learning algorithms.

Such approaches employ feature extraction mechanisms including Principal Component Analysis (PCA), Eigenfaces, Fisherfaces, Haar-like features, Histogram of Oriented Gradients (HOG), and Local Binary Patterns (LBP). The extracted features are then processed through classification algorithms capable of distinguishing facial from non-facial regions.

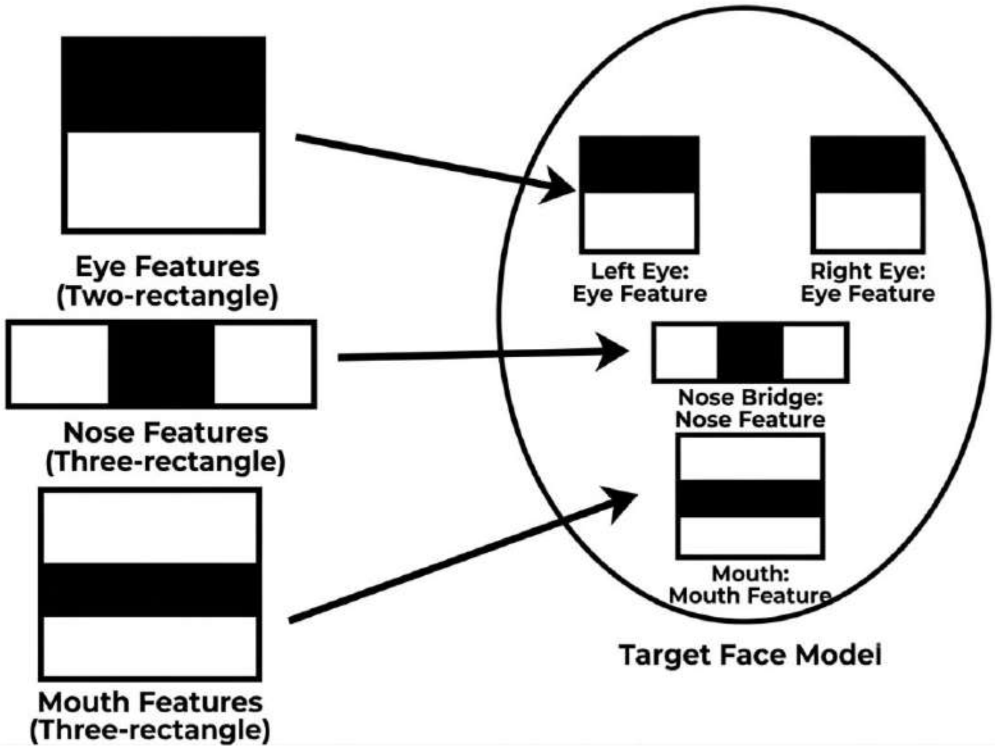


Figure 1.18: Appearance-Based Methods [36]

Appearance-based methods generally achieve higher detection accuracy than rule-based or template-based systems because they learn statistical facial distributions from real-world data. However, their performance strongly depends on dataset quality, feature selection strategies, and classifier design [27] [28].

1.4.10 Neural Networks in Face Detection

Artificial Neural Networks (ANNs) have played a transformative role in face detection and recognition research. Inspired by the structure of biological neurons, neural networks consist of interconnected computational units capable of learning complex nonlinear relationships from training data.

Early neural network-based face detectors employed shallow multilayer perceptrons trained on grayscale facial images. These systems demonstrated improved adaptability compared with handcrafted methods because they could automatically learn discriminative patterns rather than relying exclusively on predefined rules.

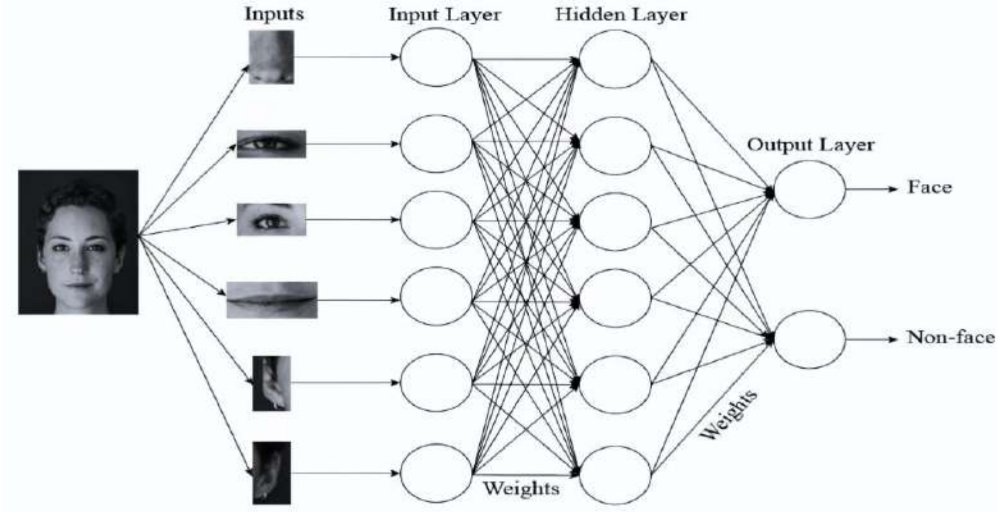


Figure 1.19: Neural Network-Based Face Detection Architecture [37]

The emergence of deep neural networks significantly revolutionized face analysis. Deep architectures, particularly Convolutional Neural Networks (CNNs), introduced hierarchical feature learning mechanisms capable of extracting increasingly abstract representations from raw images. Lower layers capture edges and textures, while deeper layers encode high-level semantic facial structures.

Neural network-based approaches offer remarkable robustness against illumination changes, facial expressions, pose variations, and partial occlusions. However, they require extensive training datasets, high computational power, and substantial memory resources. Despite these limitations, neural networks currently dominate state-of-the-art face detection systems due to their exceptional accuracy and scalability [29] [30].

1.4.11 Support Vector Machine (SVM)

Support Vector Machines (SVMs) are supervised machine learning algorithms widely utilized in face detection and classification tasks. SVMs operate by identifying an optimal hyperplane that maximizes the separation margin between different classes—in this case, facial and non-facial image regions.

In face detection systems, SVMs are frequently combined with feature extraction techniques such as HOG or LBP. The extracted features are transformed into high-dimensional vectors, allowing the SVM classifier to construct robust decision boundaries.

One of the principal strengths of SVMs lies in their excellent generalization capability, particularly when handling high-dimensional feature spaces. Kernel functions further enable nonlinear classification by projecting data into transformed spaces where separation becomes feasible.

Nevertheless, SVM performance depends heavily on parameter optimization and feature engineering quality. Moreover, training complexity increases significantly with large-scale datasets. Although deep learning has reduced the dominance of SVMs in recent years, they remain highly effective in lightweight real-time systems and embedded Computer Vision applications [28] [27].

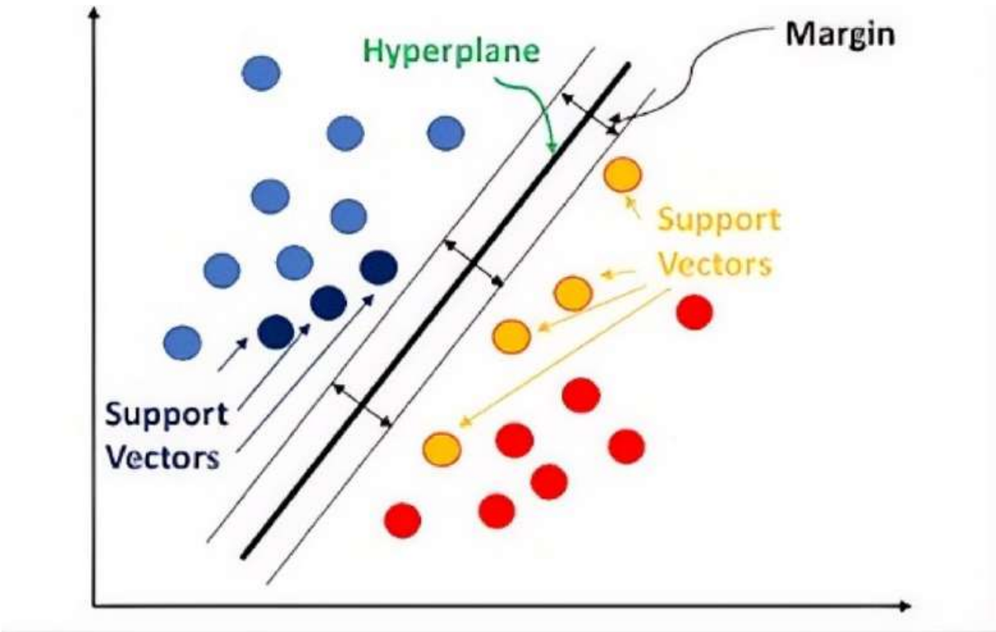


Figure 1.20: Support Vector Machine for Face Detection [38]

1.4.12 The AdaBoost-Based Face Detector

The AdaBoost-based face detector, introduced by Viola and Jones, represents one of the most influential breakthroughs in real-time face detection. This framework combines Haar-like feature extraction, integral image representation, AdaBoost learning, and cascade classification to achieve high-speed facial localization.

The AdaBoost algorithm selects the most discriminative Haar features from a large feature pool and constructs a strong classifier through iterative weighted learning. The cascade architecture further improves efficiency by quickly rejecting non-face regions during early processing stages, thereby minimizing unnecessary computations.

The Viola–Jones detector became highly popular due to its ability to perform real-time face detection on limited hardware resources. It remains widely implemented in embedded systems, robotics, surveillance platforms, and educational projects because of its simplicity and computational efficiency.

However, the method exhibits reduced robustness under non-frontal poses, severe illumination variations, and occlusions. Compared with modern CNN-based detectors, its accuracy is relatively limited in unconstrained environments. Despite these shortcomings, the AdaBoost framework remains historically significant and continues to serve as a foundational benchmark in Computer Vision research [27].

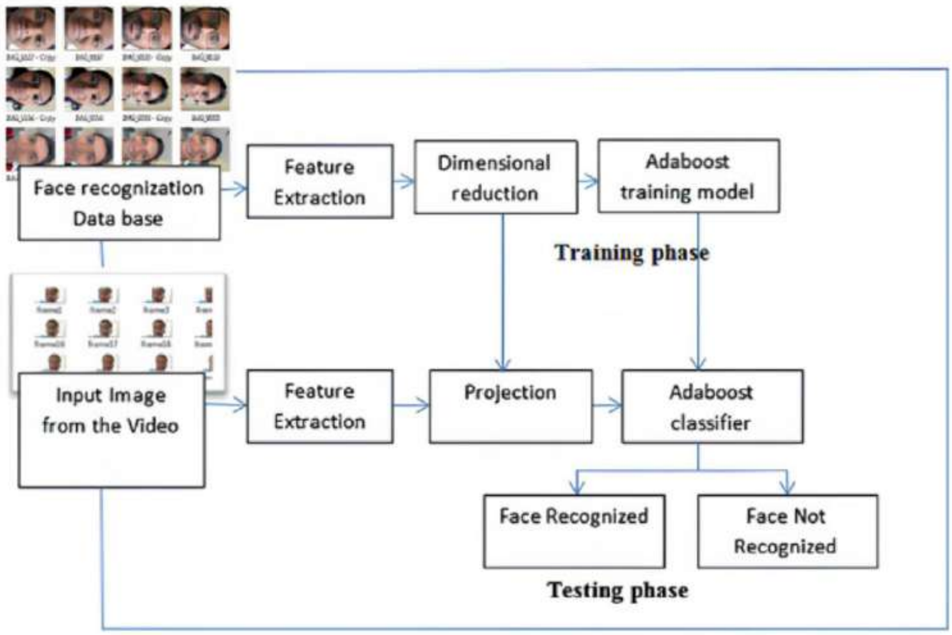


Figure 1.21: AdaBoost-Based Face Recognition System: Training and Testing Phases [39]

1.4.13 Advantages and Disadvantages of Face Recognition

Face recognition systems offer numerous advantages in intelligent environments. They provide non-contact biometric authentication, enabling convenient and hygienic user verification without requiring physical interaction. Face recognition also facilitates automation in surveillance, attendance systems, access control, and personalized user experiences. Furthermore, facial biometrics are difficult to forge compared with traditional passwords or identification cards.

Despite these advantages, face recognition technologies face substantial challenges. Recognition accuracy may degrade due to illumination changes, aging effects, pose variations, facial accessories, and occlusions. Ethical concerns related to privacy, surveillance misuse, and unauthorized data collection have also become increasingly prominent. Moreover, deep learning-based recognition systems demand large annotated datasets and powerful computational infrastructure, which may limit deployment in resource-constrained environments [40] [41].

1.4.14 Deep Learning in Face Detection and Recognition

Deep learning has fundamentally transformed the landscape of face detection and recognition. Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and transformer-based architectures now dominate state-of-the-art facial analysis systems.

Modern deep learning models such as MTCNN, RetinaFace, FaceNet, ArcFace, and DeepFace achieve remarkable performance by learning highly discriminative facial embeddings from massive datasets. These systems can operate effectively under unconstrained real-world conditions characterized by varying illumination, facial expressions, occlusions, and background complexity.

Deep learning frameworks further enable end-to-end optimization, eliminating the need for manual feature engineering. Transfer learning and pretrained models have also accelerated deployment in practical applications.

However, deep learning approaches remain computationally intensive and often require GPUs or specialized AI accelerators for real-time inference. Additionally, training large-scale models introduces challenges related to data privacy, model bias, and energy consumption. Nevertheless, deep learning remains the dominant paradigm in contemporary face analysis research and industrial deployment [29] [42] [43].

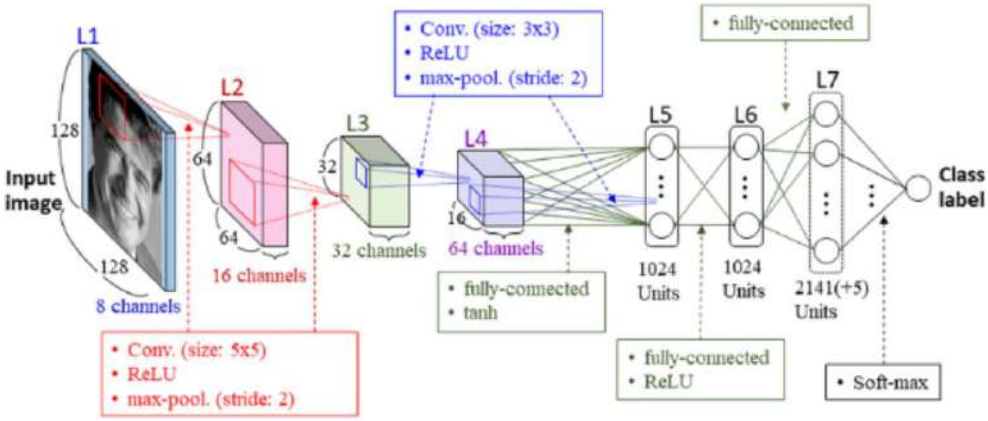


Figure 1.22: CNN Architecture for Face Recognition and Classification [44]

1.4.15 Face Recognition

Face recognition refers to the process of identifying or verifying an individual based on distinctive facial characteristics extracted from digital images or video frames. Unlike face detection, which only localizes faces, face recognition seeks to determine identity by comparing facial representations against stored templates or databases.

Face recognition systems generally operate through several sequential stages: face detection, preprocessing, feature extraction, feature matching, and identity classification. Early systems relied on geometric facial measurements and handcrafted descriptors, whereas modern approaches utilize deep neural embeddings capable of representing facial identity in high-dimensional vector spaces.

Contemporary face recognition technologies have achieved near-human accuracy in controlled conditions and are increasingly integrated into security systems, smart devices, robotics, and intelligent surveillance infrastructures. Nevertheless, recognition reliability remains influenced by environmental variability, demographic bias, and adversarial manipulation risks [42] [43] [41].

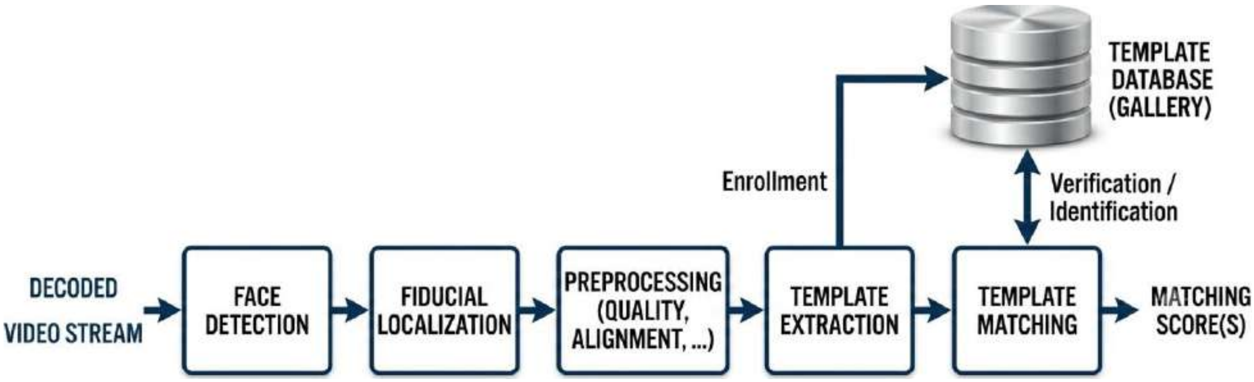


Figure 1.23: Processing Stages of a Face Recognition System [45]

1.4.16 Fundamentals of Face Recognition Systems

A conventional face recognition system comprises a sequence of interdependent computational modules designed to capture, analyze, and identify human facial features with high accuracy and robustness. The typical processing pipeline encompasses several critical stages: image acquisition, face detection, face alignment, feature extraction, feature matching, and ultimately, person identification. Each stage plays a pivotal role in ensuring the overall performance and reliability of the recognition framework.

Image Acquisition

The initial phase involves the acquisition of facial images or video sequences utilizing various imaging devices, such as digital cameras, surveillance apparatus, webcams, or embedded vision sensors. The quality and fidelity of the captured visual data profoundly influence the efficacy of all downstream recognition tasks. Key factors affecting image quality include spatial resolution, illumination conditions, camera viewpoint, focus accuracy, and motion-induced blur. In real-time intelligent systems, continuous image capture facilitates dynamic monitoring and tracking of subjects within the operational environment. Hence, optimizing image acquisition parameters is indispensable for ensuring reliable facial analysis and minimizing recognition inaccuracies [1].

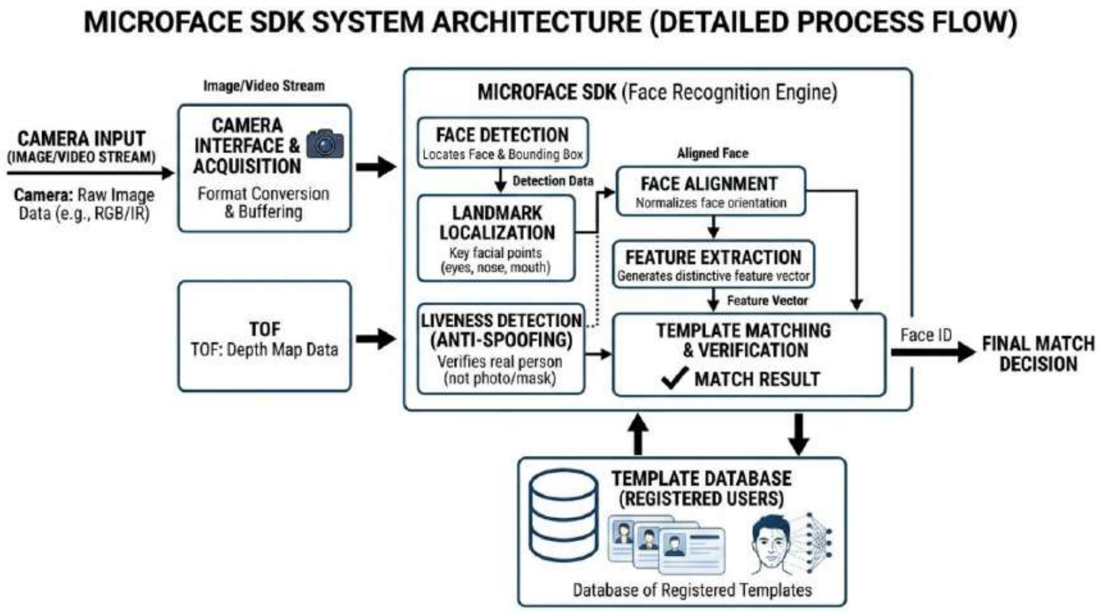


Figure 1.24: Stages of Digital Image Formation and Forensic Fingerprints [46]

Face Detection

Subsequent to image capture, the system executes face detection to localize and isolate facial regions from complex backgrounds. This module ascertains the presence of human faces within a scene and delineates their spatial extents using bounding boxes or landmark localization methodologies. Face detection algorithms may employ classical machine learning approaches such as Haar Cascade classifiers and Histogram of Oriented Gradients (HOG) descriptors, or leverage state-of-the-art deep learning architectures predominantly based on Convolutional Neural Networks (CNNs). The precision of face detection is critical, as detection errors propagate downstream and significantly degrade final identification outcomes [27] [28] [30].

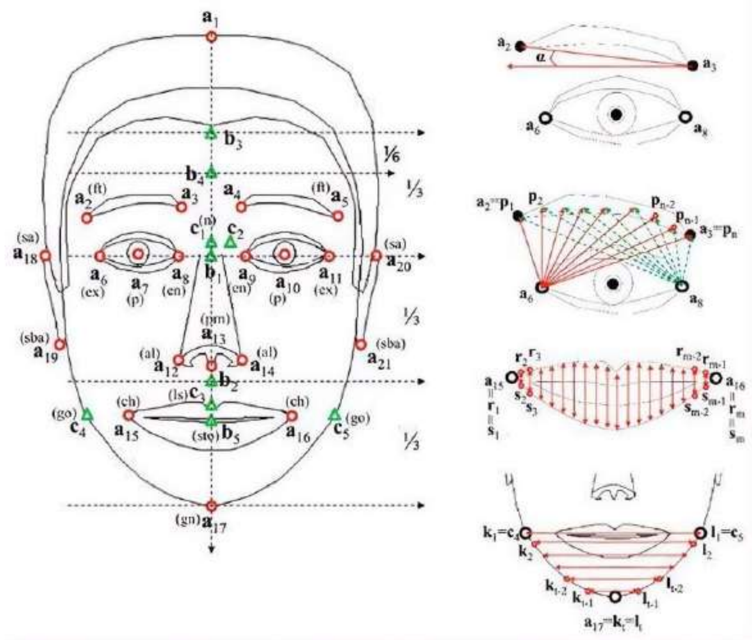


Figure 1.25: Sequence diagram of the proposed processing steps.

Face Alignment

Following detection, face alignment is performed to geometrically normalize facial images, thereby standardizing the spatial arrangement of key facial landmarks—principally the eyes, nose, and mouth. This step compensates for pose variations, head rotations, and minor facial deformations by applying geometric transformations that map the detected face to a canonical coordinate system [47] [48]. Effective alignment reduces intra-subject variability caused by differing viewpoints or facial orientations, thereby enhancing the consistency and discriminability of the features extracted in subsequent stages [49] [50]. Consequently, face alignment substantially contributes to the stability and accuracy of the recognition process [47] [48].

1.4.17 Feature Extraction

Feature extraction constitutes one of the most critical components within the face recognition pipeline, wherein the aligned facial image is processed to derive distinctive biometric descriptors that uniquely characterize an individual's identity [22] [41]. Traditional methodologies rely on hand-crafted feature representations, including Eigenfaces, Fisherfaces, Local Binary Patterns (LBP), and Histogram of Oriented Gradients (HOG) [41].In contrast, contemporary approaches predominantly employ deep learning models—specifically Convolutional Neural Networks—which autonomously learn hierarchical and discriminative facial embeddings from extensive annotated datasets [22].These learned feature vectors encapsulate identity-relevant information while exhibiting robustness against confounding factors such as illumination variability and facial expression changes [22] [41].

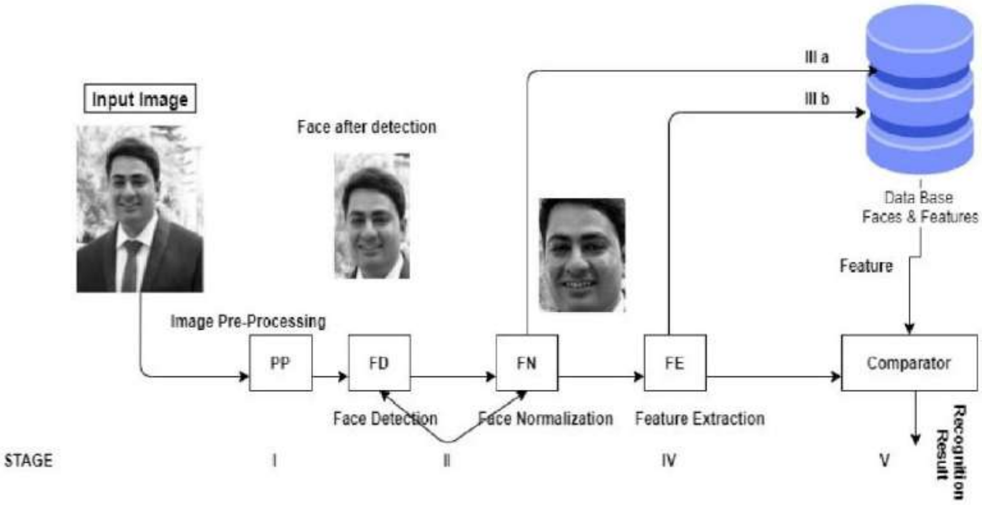


Figure 1.26: Face Recognition Process Workflow [25]

1.4.18 Feature Matching

Upon extraction, the facial feature vector is compared against a repository of stored templates within a user database during the feature matching stage [41]. This process involves computing similarity scores between the query feature vector and database entries using mathematical distance metrics such as Euclidean distance, cosine similarity, or probabilistic matching frameworks [41] [40].

The user database functions as the reference corpus of known identities against which unknown facial inputs are verified or identified [41].

The efficiency and accuracy of matching algorithms are paramount, particularly in large-scale biometric systems that require rapid comparisons across extensive datasets [41] [40].

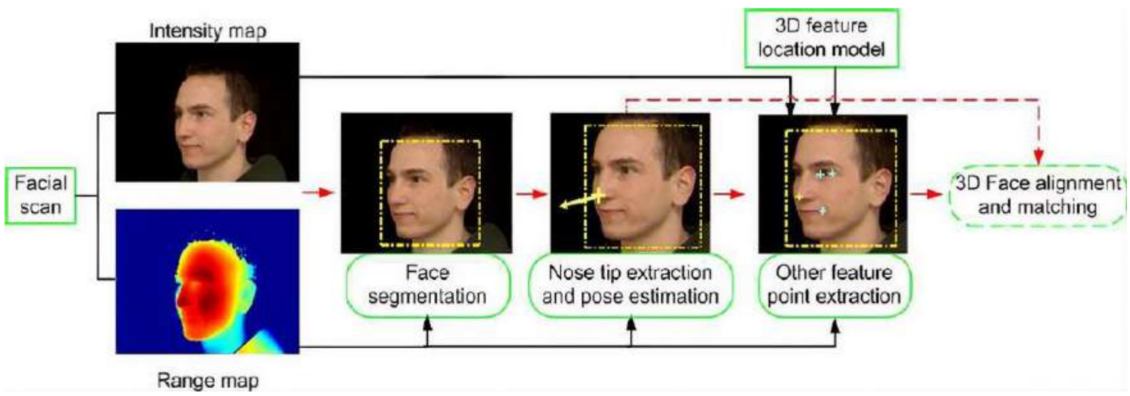


Figure 1.27: 3D Face Recognition Workflow [51]

1.4.19 Person Identification

The concluding stage is person identification, where the system assigns an identity to the detected individual based on the similarity metrics derived from feature matching [41]. A match is confirmed if the similarity score surpasses a predetermined threshold; otherwise, the individual is labeled as unknown [41]. Identification systems may operate under two principal paradigms: verification mode, entailing one-to-one comparisons, or identification mode, involving one-to-many searches [41] [48]. Advanced artificial intelligence frameworks often incorporate adaptive thresholding mechanisms, probabilistic inference models, and multimodal biometric fusion techniques to enhance decision confidence and mitigate errors such as false acceptances and false rejections [41]. The overall performance of a face recognition system is contingent upon the robustness and seamless integration of all constituent stages within the processing pipeline [41] [48]. Contemporary intelligent systems frequently augment face recognition with complementary modalities, including real-time visual tracking, speech analysis, behavioral pattern recognition, and multimodal sensor fusion [48]. Such integrative approaches facilitate augmented contextual awareness and operational reliability in complex, dynamic environments, exemplified by smart surveillance infrastructures, autonomous robotic platforms, and sophisticated human–computer interaction systems [41] [48]..

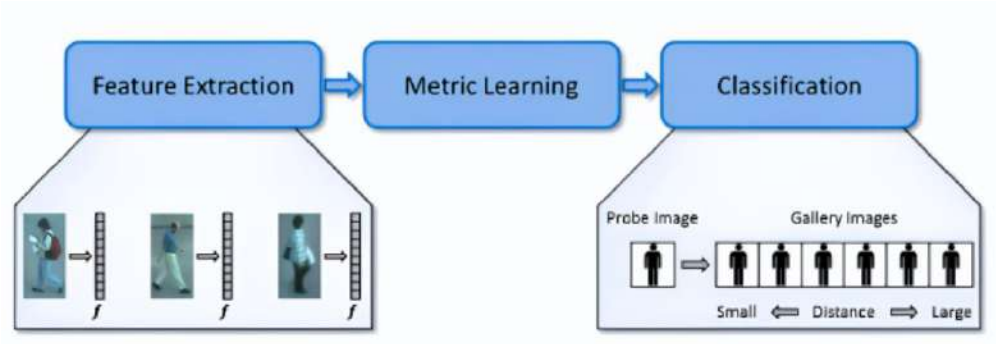


Figure 1.28: Person Re-Identification Framework [52]

1.4.20 Adopted Approach

In this thesis, a lightweight face recognition methodology has been adopted to achieve an optimal balance among recognition accuracy, computational efficiency, and memory footprint [41] [53]. This strategic selection is driven by the imperative to ensure that the system can operate effectively within real-time constraints [41].

Face analysis systems inherently exhibit sensitivity to a variety of environmental and subject-specific factors that can adversely affect their accuracy and robustness [41] [40].

Illumination variations represent a primary challenge, as changes in lighting conditions can substantially alter the visual appearance of facial features [41] [48].

Moreover, intra-class variability arising from facial expressions and natural aging processes further complicates the recognition task by introducing temporal inconsistencies in facial appearance [41] [40].

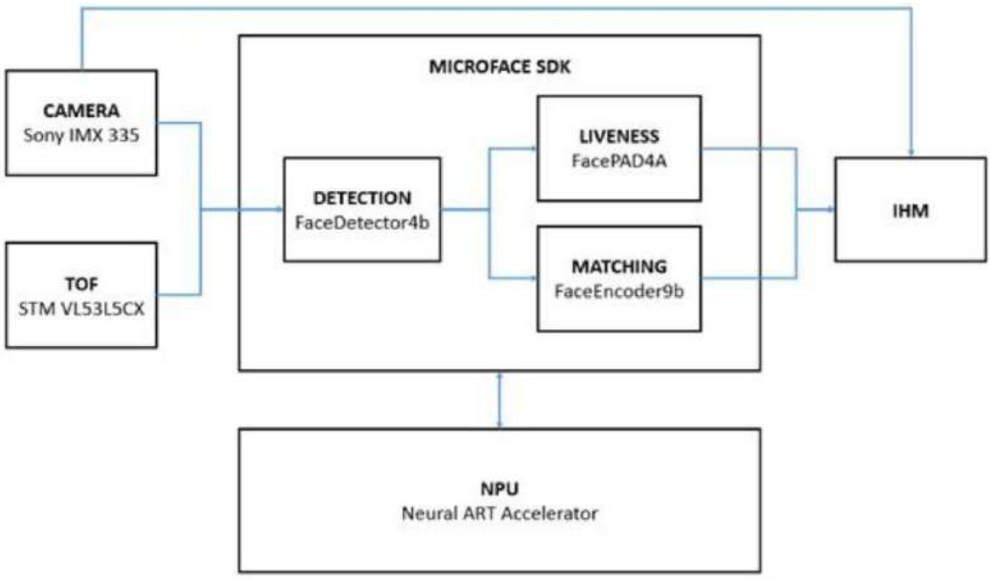


Figure 1.29: MICROFACE SDK Architecture for Face Recognition and Liveness Detection [54]

1.4.21 Face Pose Estimation

Face pose estimation is a critical task in contemporary computer vision that involves determining the three-dimensional orientation of a human head relative to the camera coordinate system. This orientation is commonly quantified using three Euler angles—yaw, pitch, and roll—which respectively correspond to horizontal rotation, vertical tilt, and in-plane rotation of the head. These parameters succinctly characterize head posture and are extensively applied in intelligent tracking systems, human–robot interaction, and augmented reality contexts **zhao23 szeliski22**.

Recent progress from 2023 to 2026 has markedly enhanced the robustness and accuracy of face pose estimation under challenging real-world conditions such as occlusions, varying illumination, and rapid or extreme head movements. These advancements primarily stem from the adoption of deep learning architectures and hybrid models that integrate data-driven feature extraction with geometric constraints, enabling more reliable pose inference in diverse scenarios **zhao23 wang25**.

Methodological Approaches

Several methodological approaches underpin face pose estimation. Landmark-based techniques detect key facial points (e.g., eyes, nose, mouth corners) and compute geometric relationships among these landmarks to estimate head orientation. These methods are computationally efficient and suitable for real-time applications but are sensitive to occlusion and inaccuracies in landmark detection **zhao23**. Geometric methods, particularly Perspective-n-Point (PnP) algorithms, solve for pose parameters by establishing correspondences between 2D landmarks and a predefined 3D facial model. They offer a balanced trade-off between computational cost and accuracy but depend heavily on the precision of landmark localization **szeliski22**. Deep learning-based approaches directly regress head pose angles from raw image data using convolutional neural networks (CNNs) or transformer models. While these end-to-end models provide high robustness and generalization across varying conditions, they demand large annotated datasets and significant computational resources, which can restrict deployment on resource-constrained platforms **wang25** | Dosovitskiy et al., 2024 |.

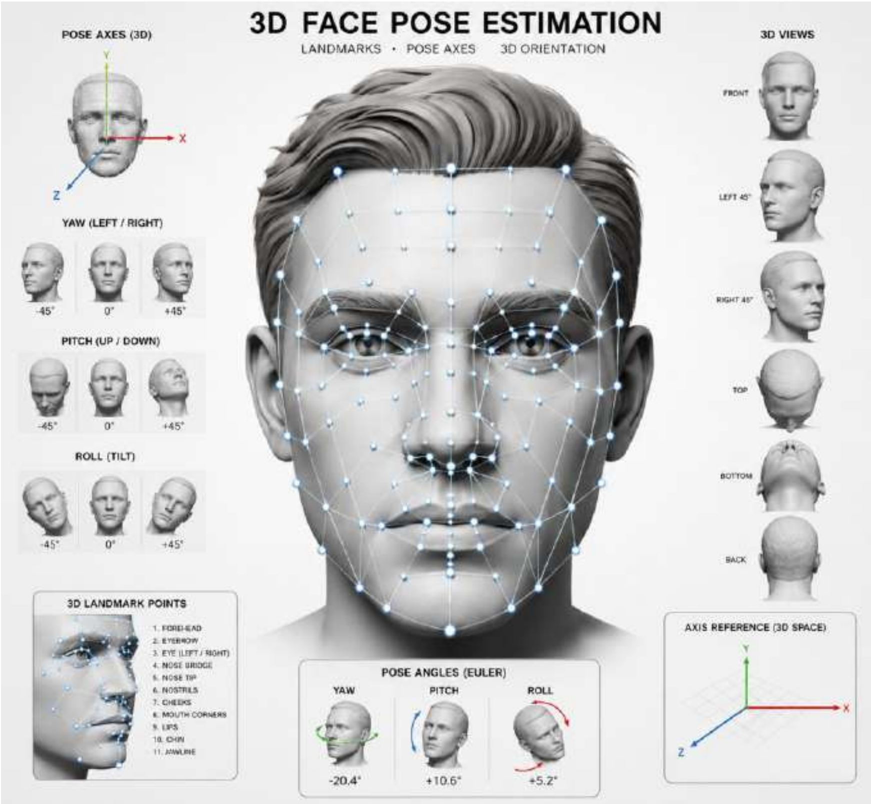


Figure 1.30: 3D Face Pose Estimation and Facial Landmark Localization [34]

Comparative Analysis of Techniques

Comparative analyses reveal that landmark-based methods excel in speed but struggle with occlusions, PnP-based methods balance accuracy and efficiency contingent on landmark quality, and deep learning models achieve superior accuracy and robustness at the expense of higher computationalload **zhao23 szeliski22 wang25** | Dosovitskiy et al., 2024 |.

Technique	**Principle**	**Advantages**	**Limitations**
Landmark-Based	Facial keypoint geometry	Fast, low computational cost	Sensitive to occlusion and noise
PnP Methods	2D–3D geometric projection	Balanced accuracy and efficiency	Depends on landmark precision
Deep Learning	End-to-end regression models	High robustness and accuracy	High computational cost

Table 1.3: Comparative Analysis of Face Pose Estimation Techniques

Role in Intelligent Tracking Systems

In intelligent vision-driven tracking systems, face pose estimation is vital for maintaining precise alignment between the camera and the target. Continuous estimation of head orientation enables dynamic adjustment of pan-tilt mechanisms, ensuring the subject remains centered within the camera's field of view. Recent studies emphasize that integrating pose estimation with real-time feedback control substantially enhances tracking stability and minimizes target loss in dynamic environments, establishing pose estimation as a cornerstone in modern autonomous tracking architectures **zhao23 wang25**.

1.4.22 Proposed Contribution

This project advances beyond conventional software-only pose correction methodologies by proposing a hybrid framework that integrates a mechanical intervention to enhance system performance. Specifically, the approach employs a robotic manipulator to physically adjust the camera's orientation in real time, thereby maintaining optimal alignment with the subject's face. By offloading part of the pose correction task to the robotic arm, the system achieves improved tracking robustness and stability, while simultaneously alleviating the computational burden typically associated with purely algorithmic pose adjustments [55] [56] [57] [58].

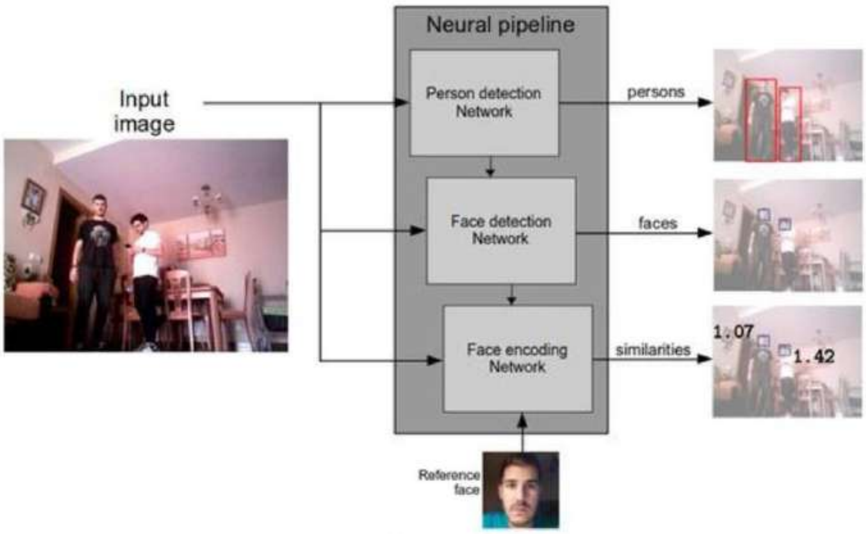


Figure 1.31: Face Recognition Neural Pipeline [59]

1.5 Intelligent Tracking Systems

1.5.1 Concept and Significance

Intelligent tracking systems are autonomous frameworks engineered to detect, monitor, and continuously follow targets in real time. They achieve precise and uninterrupted tracking by integrating advanced Computer Vision techniques, signal processing methodologies, and control system architectures [60] [1].

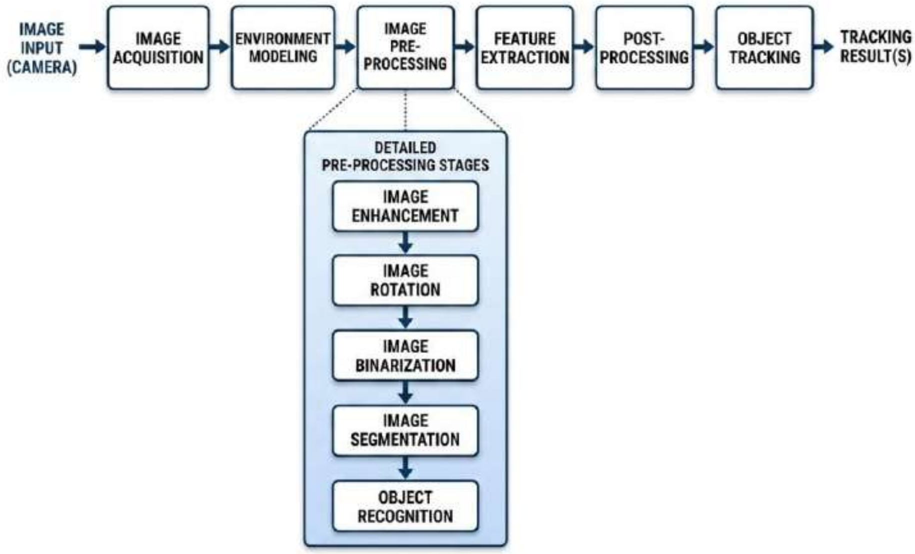


Figure 1.32: Vision Sensor Data Processing Pipeline [61]

1.5.2 Deployment in Smart Environments

The deployment of intelligent tracking systems is increasingly prevalent across diverse smart environments, including intelligent classrooms, advanced surveillance infrastructures, and human–robot interaction platforms. Within these settings, such systems facilitate enhanced automation by dynamically responding to human presence and behavioral cues **Hu2004**, **Wang2013** .

1.5.3 Rationale and Application Context

The critical importance of intelligent tracking systems has been markedly accentuated in recent years, particularly in response to global challenges such as the COVID-19 pandemic, which underscored the necessity for remote communication and automated monitoring solutions [62], [63].

Consequently, the design and development of efficient, adaptive, and robust tracking solutions have emerged as pivotal research priorities within the domains of artificial intelligence and robotics [64] [65].

1.8 Market Analysis and System Positioning

A variety of commercial solutions have been introduced to cater to automated tracking requirements, exemplified by products such as the DJI Osmo and OBSBOT Tail, which predominantly utilize vision-based algorithms to enable their tracking functionalities [71] [72].

Despite their commercial viability, these existing solutions present notable limitations. Their reliance on exclusively vision-centric tracking methodologies renders them vulnerable to common challenges such as occlusions, fluctuating illumination conditions, and significant variations in target pose [60] [73].

The system proposed herein seeks to overcome these deficiencies by adopting a multimodal integration strategy, combining face recognition, pose-aware tracking, and audio-visual fusion techniques. This comprehensive framework substantially enhances system robustness and elevates tracking precision under real-world conditions [74] [75] [76].

1.9 System Limitations and Bimodal Approach

1.9.1 System Limitations

Despite significant progress in the development of intelligent tracking systems, several inherent limitations persist. Rapid target movements can result in transient tracking failures attributable to system latency and delayed response. Additionally, fluctuations in ambient lighting conditions adversely affect the reliability and accuracy of vision-based algorithms [60] [1].

These challenges underscore the necessity for more resilient and adaptive tracking methodologies capable of maintaining reliable operation amid dynamic and unpredictable environmental conditions

[77] [78].

1.9.2 Proposed Solution

To address these shortcomings, the present work adopts a bimodal sensory fusion strategy that synergistically combines auditory and visual data streams. This multimodal integration enhances tracking robustness by leveraging the complementary strengths of each modality: for example, in scenarios where visual tracking is compromised due to occlusion or poor visibility, audio-based localization can provide directional cues to sustain target acquisition [74] [75].

Moreover, the implementation incorporates a Kalman Filter to mitigate measurement noise and enhance tracking stability. As an optimal recursive estimator, the Kalman Filter facilitates accurate state estimation of the dynamic system by smoothing motion trajectories and producing reliable predictions, all while maintaining computational efficiency suitable for real-time applications [79] [80].

1.10 Proposed System Architecture

The proposed system architecture synthesizes Computer Vision, audio signal processing, and robotic control within a cohesive, closed-loop framework. This architecture enables continuous perception-action coupling, wherein sensory inputs are processed in real time to inform and adapt actuator commands dynamically, thereby supporting responsive, autonomous tracking behavior in complex operational environments [64] [65].

1.10.1 System Workflow

The operational workflow of the proposed system is delineated through a sequence of interconnected stages, each contributing to the overall objective of robust and continuous target tracking [60] [1].

Table 1.4: System Workflow Stages

Stage	Description
Image Acquisition	Capturing visual data using a camera
Face Detection	Identifying potential faces in the scene
Face Recognition	Determining the identity of detected individuals
Target Selection	Choosing the subject of interest
Motor Control	Generating control signals for the robotic arm
Tracking	Continuously adjusting camera orientation
Fallback (Audio-based)	Switching to audio localization when visual tracking fails

This integrated, hybrid workflow capitalizes on the complementary strengths of visual and auditory modalities, enabling resilient target tracking under diverse and challenging operational scenarios [1] [65].

1.10.2 Comparative Analysis of Robotic Arms

1.11 Selection of Robotic Arm for the Proposed System

Within the framework of the proposed intelligent tracking system—centered on vision-driven perception and automated camera orientation—the selection of the robotic manipulator constitutes a pivotal architectural decision that profoundly impacts overall system efficacy. Designed to function in real-time within dynamic and unpredictable environments, the system relies on continuous visual feedback to detect, track, and sustain alignment with a moving target [1] [3].

To guarantee reliable and stable performance under these stringent conditions, the robotic platform must satisfy rigorous operational criteria. These include the capability to execute smooth, continuous motions devoid of abrupt transitions, low-latency responsiveness to real-time control commands, and coordinated multi-axis movement within a three-dimensional workspace [2] [4].

Considering these technical and functional imperatives, articulated robotic arms emerge as the most suitable mechanical configuration for the proposed system. Characterized by a serial chain of rigid links interconnected via rotary joints, articulated arms afford multiple degrees of freedom that closely emulate the kinematic behavior of the human upper limb [2] [18].

Technically, several key advantages substantiate the selection of articulated arms. First, their elevated degrees of freedom significantly enhance system manipulability, facilitating precise real-time control over camera orientation, position, and angular alignment. Second, the rotary joint configuration supports smooth trajectory generation, which is critical to maintaining stable visual tracking. Third, articulated arms exhibit robust adaptability to dynamic contexts, as their kinematic redundancy provides multiple feasible joint configurations [4] [18].

In summary, robotic manipulators serve a central role in contemporary intelligent systems requiring precise interaction with dynamic, partially unpredictable environments. While various robotic architectures possess distinct advantages and limitations, articulated robotic arms offer an optimal synthesis of flexibility, accuracy, and controllability for vision-based real-time tracking systems [1] [3].

1.11.1 Adopted Mechanism in This Project

In addition to the broader robotic architecture integrated within this system, a pan-tilt rotational mechanism has been strategically selected as a pragmatic and effective approach for camera orientation control. This mechanism provides two rotational degrees of freedom—pan (yaw rotation) and tilt (pitch rotation)—which are widely used in vision-based tracking systems to adjust camera orientation while keeping a fixed base position [1] [4].
The decision to implement a pan-tilt mechanism is primarily motivated by its inherent structural simplicity, cost-effectiveness, and seamless compatibility with embedded and real-time processing platforms. Compared to fully articulated robotic manipulators, pan-tilt systems significantly reduce mechanical complexity and computational requirements by limiting motion to two controlled axes [4] [2].

Despite its reduced degrees of freedom, the pan-tilt architecture demonstrates strong performance when combined with Computer Vision algorithms and feedback control strategies. Its simplified kinematic model allows efficient real-time control implementation, which is essential for latency-sensitive tracking applications [1] | Hartley & Zisserman, 2004 | [3].

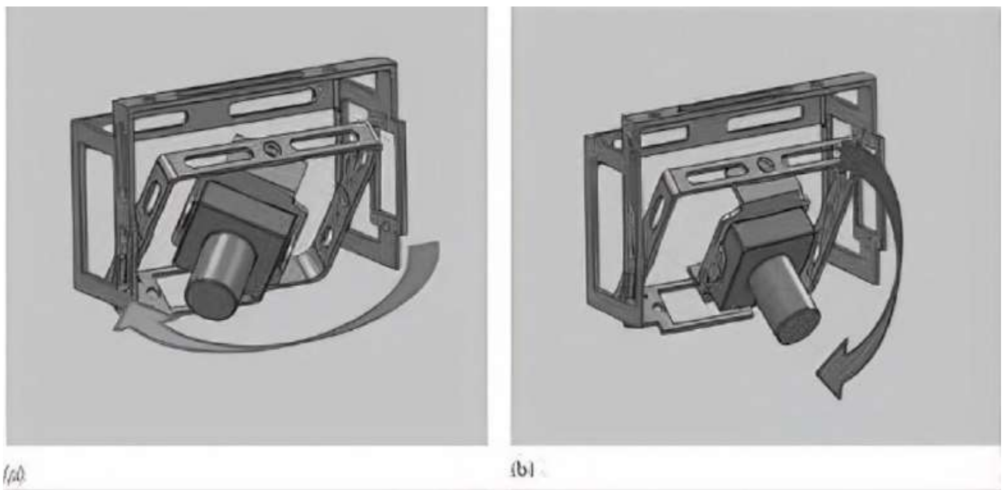


Figure 1.38: Pan-Tilt Camera Mechanism

Source: Pan-Tilt Camera Mechanism

1.12 Chapter Conclusion

This chapter provides a comprehensive overview of intelligent tracking systems that integrate computer vision, audio processing, and robotic control to enable autonomous real-time tracking in dynamic environments. It outlines the evolution of AI from theory to practical systems capable of perceiving and interacting with complex surroundings **zhao23**.
Robotic arms, especially articulated types with multiple degrees of freedom, are emphasized for their flexibility and precision in aligning sensors with moving targets **correll22**.
Computer vision is described as a hierarchical process powered by deep learning models, facing challenges such as variable lighting and computational constraints **szeliski22**.
Face detection and recognition have evolved from traditional handcrafted methods to advanced deep learning architectures that improve robustness under diverse conditions **zhao23 szeliski22**.
Audio processing and sound source localization methods like TDOA complement vision by accurately identifying sound source positions, enhancing system reliability **chen24**. Multimodal fusion of audio-visual data, supported by filtering techniques, significantly improves tracking performance when visual input is limited. The proposed system architecture combines these modalities with robotic pan-tilt control to maintain precise camera orientation and robust autonomous tracking **chen24 szeliski22**.
The chapter establishes a solid foundation for developing advanced intelligent systems that unify sensory perception and robotic actuation for reliable operation in real-world scenarios **zhao23 correll22 chen24 szeliski22**.

PRACTICAL PART

Implementation, Experiments, and Results

Chapter 2

Practical Implementation of an AI-Driven Face Tracking Robotic Arm

2.1 Introduction to the Practical Implementation

This chapter presents the complete practical implementation of the TeleGimbal system: a smart rotational robotic arm equipped with an intelligent face-tracking mechanism based on Deep Neural Networks (DNNs). The primary objective of this chapter is to bridge the gap between the theoretical concepts discussed in previous chapters and their real-world application. The TeleGimbal system is designed to autonomously detect, recognize, and track a human face in real time, translating visual information from a USB webcam into precise physical movements of a two-axis robotic arm (Pan for horizontal rotation and Tilt for vertical inclination) [111].
The chapter is organized into three main sections corresponding to the system's architectural layers. First, the software layer is presented, encompassing the graphical user interface (GUI) developed using Java Swing, the computer vision pipeline built with OpenCV's DNN module, and the serial communication middleware using the `jSerialComm` library. The software layer is responsible for real-time face detection using the YuNet ONNX model, face recognition using the SFace ONNX model, and the Bang-Bang velocity controller that converts pixel errors into directional motor commands [112][113]

Second, the hardware layer is documented in detail, including the central microcontroller (Arduino Uno), the dual NEMA 17 stepper motors for the Pan and Tilt axes, the A4988 motor drivers with 1/16 microstepping capability, the 12V DC power supply, and the 3D-printed mechanical support structure. The principles governing the operation of these actuators and drivers are well established in the literature on precision motion control [114][115].

Third, the integration and testing phase is presented, where the complete system is evaluated against key performance metrics including frame rate, tracking latency, detection and recognition confidence, motor temperature, and angular displacement measurement [116].

The practical challenges encountered during implementation are systematically documented, including acoustic resonance and mechanical jitter, kinetic overshoot (the pendulum effect), AI-induced bounding box jitter, runaway motor behavior following target loss, and VREF current calibration for thermal management. For each challenge, a detailed solution is provided, demonstrating engineering rigor and a structured troubleshooting methodology [117] [118] .

The chapter concludes with a comprehensive presentation of the testing results, validating that the system achieves frame rates between 27.75 and 29.79 FPS, recognition confidence scores exceeding the 0.7 threshold, and smooth vibration-free tracking across a 30 pan range and a 20 tilt range.

Unlike traditional face-tracking systems that rely on absolute position control or require expensive GPU-based computation, the TeleGimbal system demonstrates that a decoupled architecture—where computationally intensive AI inference executes on a standard CPU while real-time motor control is delegated to a dedicated microcontroller—is both feasible and highly effective for real-time applications [119] [120]. This implementation contributes to the growing field of accessible AI-driven robotics by providing a low-cost, open-source solution for intelligent face tracking, aligning with recent advances in embedded systems and service robotics [121].

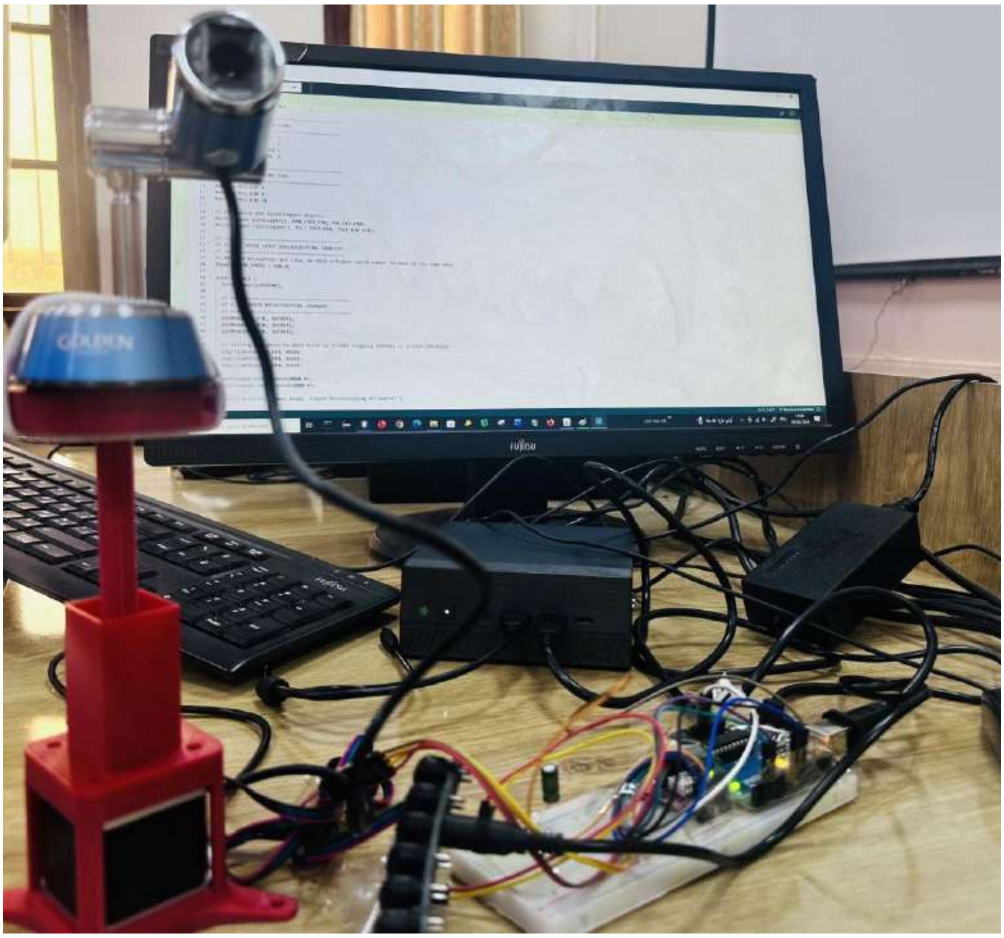


Figure 2.1: TeleGimbal Hardware Prototype

The physical prototype of the TeleGimbal system consists of a USB camera, an Arduino Uno microcontroller, two NEMA 17 stepper motors, A4988 motor drivers, and a custom mechanical support structure. These components work together to enable real-time face tracking through coordinated Pan and Tilt movements.

2.2 Global Software Architecture

To achieve real-time tracking performance at a minimum of 30 frames per second (FPS), the software was engineered using a strict multithreaded architecture. Running the entire system on a single thread would cause the Graphical User Interface (GUI) to freeze every time a frame was processed or a serial command was dispatched [116]. Therefore, the software is divided into three asynchronous domains:

1. **The Main GUI Thread:** Handles user inputs, displays the live video feed, manages the biometric face database, and visualizes tracking status (e.g., `"TARGET LOCKED"`).

2. **The Vision Processing Thread:** A dedicated background thread that continuously captures webcam frames, executes DNN inferences (YuNet for detection and SFace for recognition), and calculates pixel-based tracking errors.

3. **The Communication Middleware:** A non-blocking serial dispatch system using the `jSerialComm` library that pushes velocity commands to the microcontroller without waiting for hardware acknowledgments [122].

This decoupled design prevents any single component from becoming a bottleneck, a key requirement for real-time cyber-physical systems | Liu et al., 2024 |.

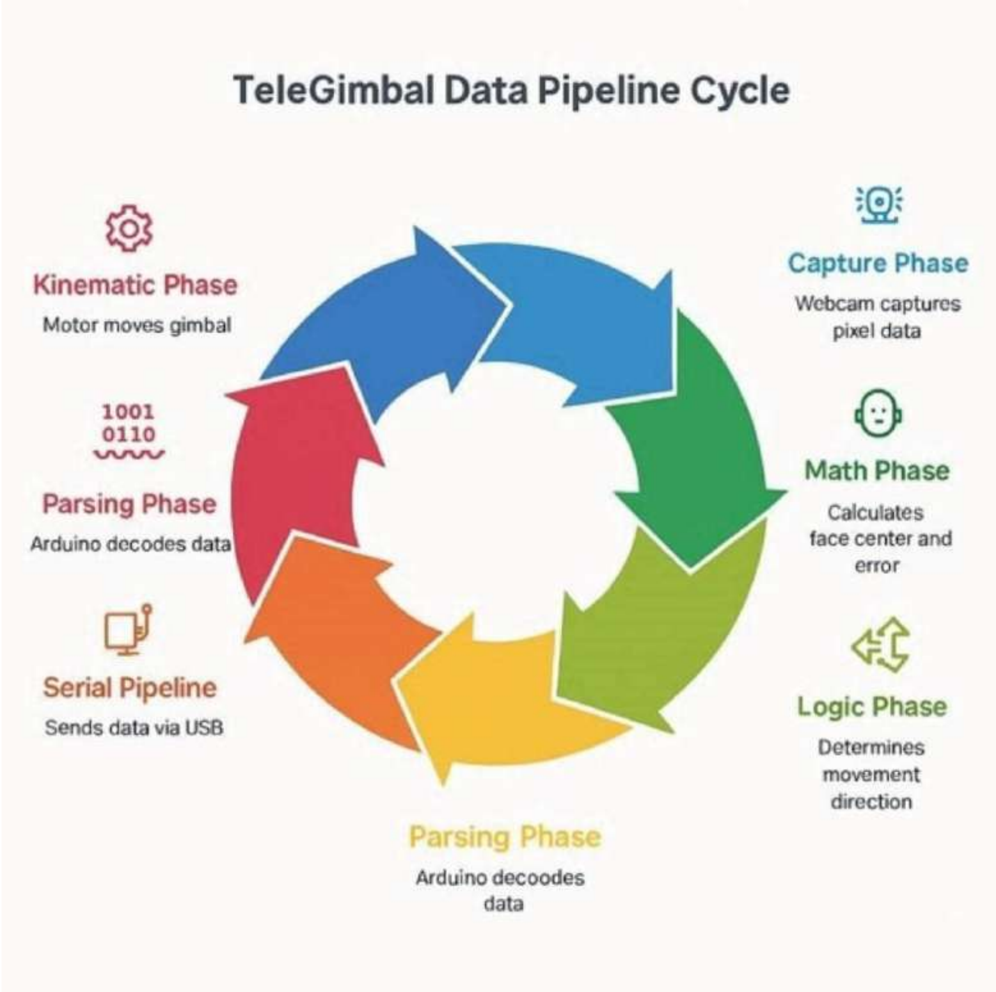


Figure 2.2: TeleGimbal Data Processing Pipeline

Figure2.2 illustrates the complete TeleGimbal data pipeline, from image acquisition and face processing to command generation and motor actuation. It highlights the transition from AI-based visual perception to real-time robotic motion control.

2.3 The Perception Layer: Computer Vision Implementation

The vision layer is built using the OpenCV (Open Source Computer Vision Library) with Java bindings. Instead of using legacy algorithms such as Haar Cascades—which are prone to false positives and highly sensitive to lighting variations—the system leverages modern Deep Learning models running through OpenCV's DNN (Deep Neural Network) module [123].
Recent comparative studies have shown that DNN-based detectors outperform traditional methods by a significant margin in unconstrained environments [124].

2.3.1 Face Detection: YuNet

The system utilizes the YuNet ONNX model for face detection (`face_detection_yunet_2023mar.onnx`). YuNet is a highly optimized, lightweight

Convolutional Neural Network (CNN) specifically designed for real-time applications [112]. The model outputs a bounding box defined by four numbers: `(X, Y, Width, Height)`. The detection pipeline is implemented as follows:

```java
// Load the YuNet model from the specified path
Net faceDetector = Dnn.readNetFromONNX("F:/DNN_Face/
    face_detection_yunet_2023mar.onnx");

// Prepare the input blob
Mat blob = Dnn.blobFromImage(frame, 1.0, new Size(320, 320));
faceDetector.setInput(blob);

// Perform detection
Mat detections = faceDetector.forward();
```

Figure 3.3: YuNet face detection implementation

2.3.2 Face Recognition: SFace

When operating in *"Specific Target"* mode, the system passes the detected face bounding box from YuNet to the SFace ONNX model (`face_recognition_sface_2021dec.onnx`), which extracts a 128-dimensional biometric feature vector (embedding) [113] This vector is compared against an in-memory `HashMap` (the face database) using Cosine Distance matching. A recognition confidence threshold of 0.7 was empirically determined to balance precision and recall [125].

```java
// Load the SFace model
Net faceRecognizer = Dnn.readNetFromONNX("F:/DNN_Face/
    face_recognition_sface_2021dec.onnx");

// Extract embedding from detected face region
Mat faceRegion = new Mat(frame, boundingBox);
Mat blob = Dnn.blobFromImage(faceRegion, 1.0, new Size(112, 112));
faceRecognizer.setInput(blob);
Mat embedding = faceRecognizer.forward();
```

Figure 3.4: SFace face recognition implementation.

2.4 Graphical User Interface (GUI) Design

The graphical user interface was developed using Java Swing to provide intuitive control over the tracking system [126]. The interface displays real-time information including communication port configuration (COM10), file paths to YuNet and SFace ONNX models, live video stream with options to show confidence scores and facial landmarks, tracking status (Pan and Tilt motor states), target information (detected faces count, detection confidence at 0.6 threshold), recognized person name with confidence score, tolerance sliders (X:50, Y:50 pixels), and database management buttons.

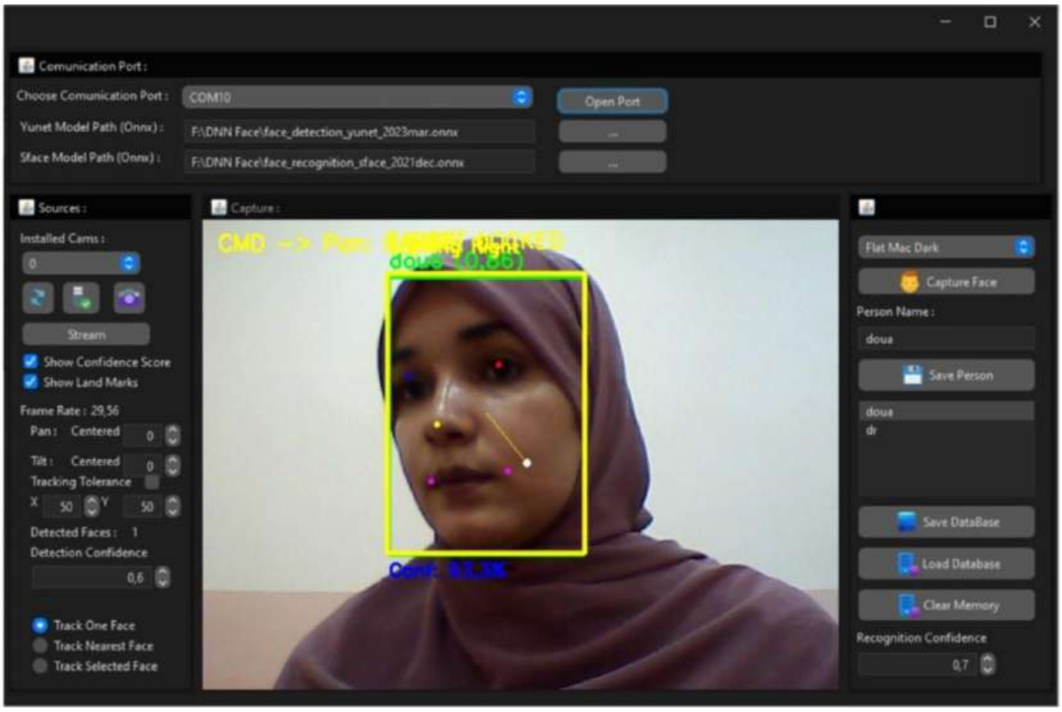


Figure 2.5: Main graphical user interface during active face tracking.

The interface displays communication port COM10, YuNet and SFace model paths, frame rate 29.76 FPS, Pan status Centered with Tilt UP, command CMD ! Pan: 0 j Tilt: 1, target "doua" locked with recognition confidence 0.79, and tracking tolerance X:50, Y:50 pixels.

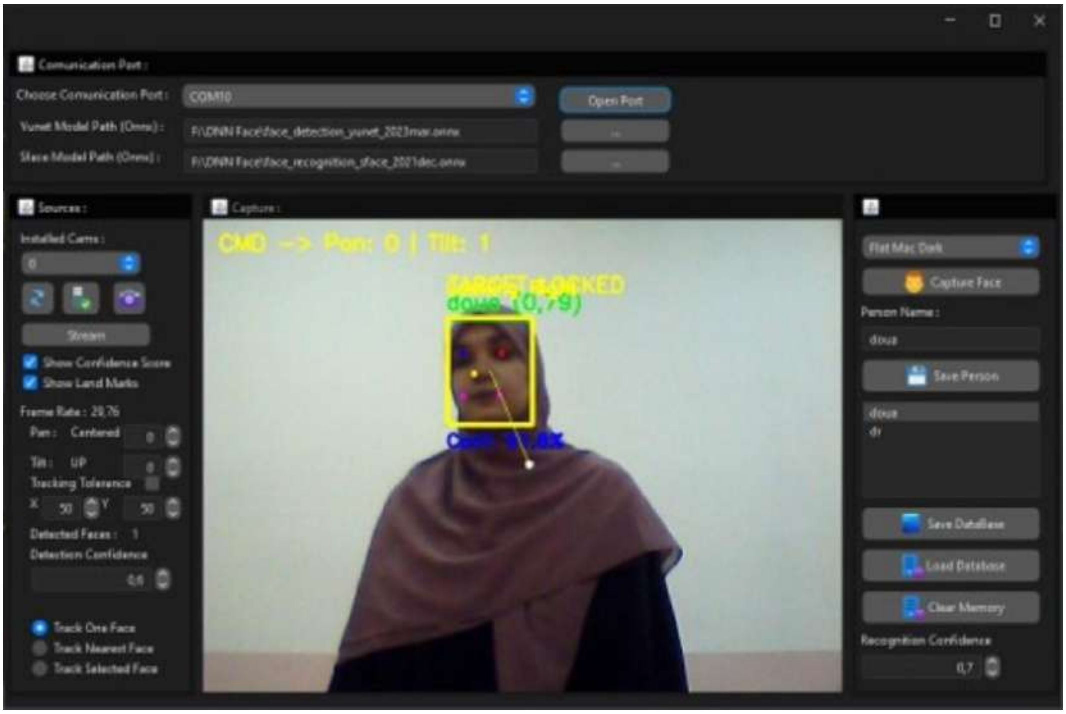


Figure 2.6: Graphical interface showing pan movement during target tracking.

The interface shows frame rate 29.79 FPS, Pan status LEFT with Tilt UP, command CMD ! Pan: 1 j Tilt: 1, and tilt angle displayed at 92:14.

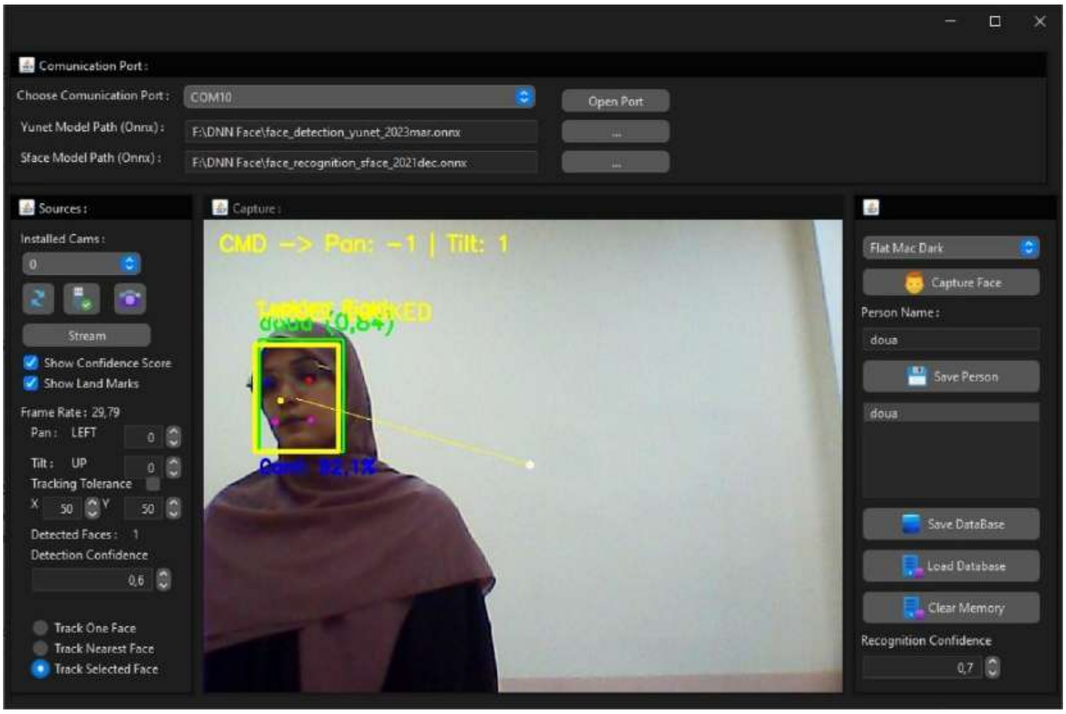


Figure 2.7: Simultaneous Pan-Tilt Tracking

The GUI during diagonal tracking movement where both Pan and Tilt motors are active simultaneously. The command `CMD -> Pan: -1 | Tilt: -1` indicates Pan moving left and Tilt moving down, with target "doua" recognized at a confidence score of 0.87.

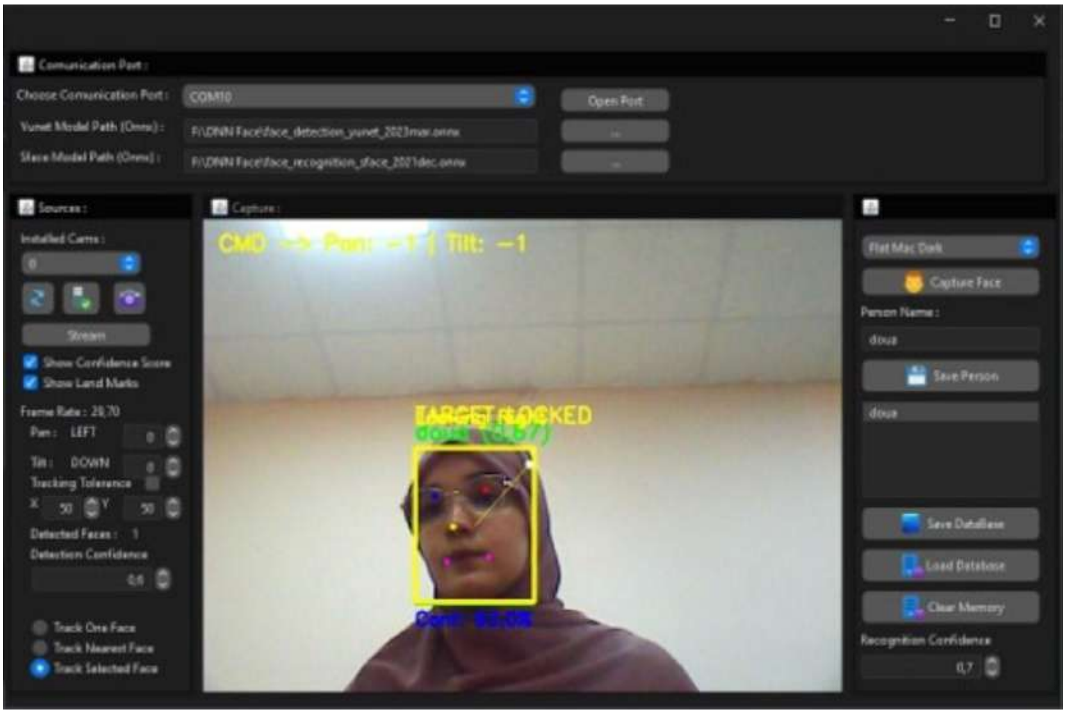


Figure 2.8: Graphical interface showing simultaneous pan and tilt tracking.

The GUI during successful target recognition showing "doua" with confidence 0.86, frame rate 29.56 FPS, and Tilt status Centered.

3.5 Tracking Logic and Decision Making

After the perception layer identifies the target, the system must calculate how to actuate the hardware. The transition from pixel coordinates to physical motion required a highly optimized control algorithm. Initial experiments using absolute position commands resulted in severe latency and overshoot, a common issue in vision-based servoing systems [127].

3.5.1 Cartesian Error Mapping

The algorithm calculates the exact pixel center of the camera frame (`Center_Frame`) and the target's bounding box (`Center_Target`):

```
int frameCenterX = frame.width() / 2;
int frameCenterY = frame.height() / 2;
int faceCenterX = faceRect.x + faceRect.width / 2;
int faceCenterY = faceRect.y + faceRect.height / 2;

int errorX = faceCenterX - frameCenterX; // Horizontal error
int errorY = frameCenterY - faceCenterY; // Vertical error
```

Figure 2.9: Cartesian error mapping.

3.5.2 The Tolerance Box (Deadzone)

A programmable tolerance box of 50 pixels in both X and Y directions was implemented. The selected 50-pixel tolerance provides an optimal balance between stability and tracking accuracy based on empirical testing [128].

```
if (Math.abs(errorX) < TOLERANCE_X && Math.abs(errorY) < TOLERANCE_Y
    ) {
panDirection = 0; // Stop Pan motor
tiltDirection = 0; // Stop Tilt motor
} else {
panDirection = (errorX > 0) ? 1 : -1;
tiltDirection = (errorY > 0) ? 1 : -1;
}
```

Figure 2.10: Tolerance box implementation.

3.5.3 Motor Command Generation

As shown in Figure 3.7, the command `P:-1, Y:-1` indicates simultaneous diagonal movement, proving the system can handle complex two-dimensional target motion.

Command Format	Meaning
P:0,Y:0	Stop both motors
P:1,Y:0	Pan right
P:-1,Y:0	Pan left
P:0,Y:1	Tilt up
P:0,Y:-1	Tilt down
P:1,Y:1	Diagonal movement: right and up
P:-1,Y:-1	Diagonal movement: left and down

Table 3.1: Motor Command Formats and Their Corresponding Meanings

3.6 Serial Communication Middleware

The bridge between the Java environment (running on a standard PC) and the Arduino microcontroller is handled by the `jSerialComm` library [129]. The system communicates through COM10 at a baud rate of 115,200.

```
SerialPort comPort = SerialPort.getCommPort("COM10");
comPort.setBaudRate(115200);
comPort.openPort();

// Non-blocking write
comPort.writeBytes("P:1,Y:0\n".getBytes(), 7);
```

Figure 2.11: jSerialComm implementation.

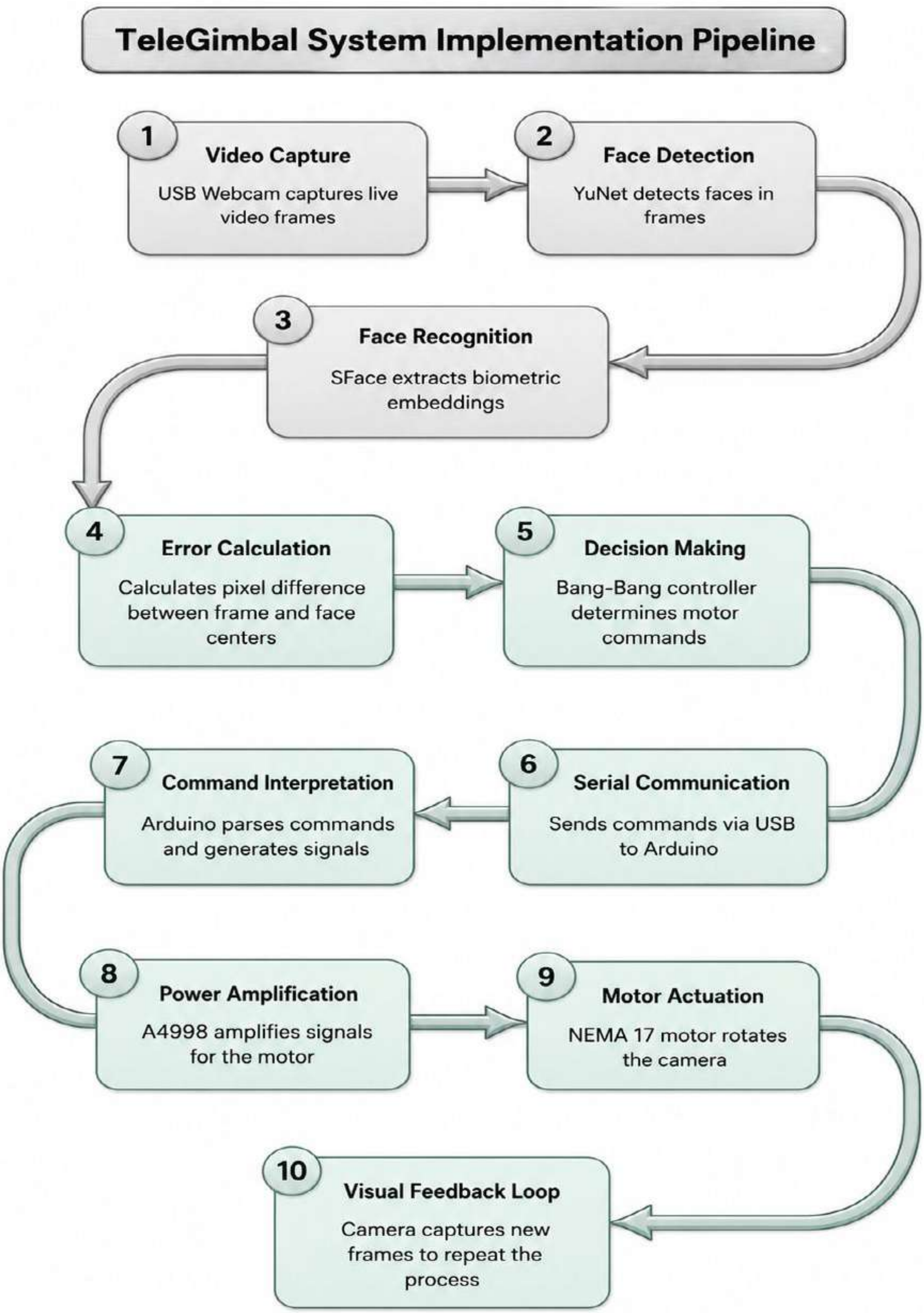


figure 2.12:illustrates the TeleGimbal system workflow.

2.7 Global Hardware Architecture

While the software layer acts as the cognitive brain, the hardware layer serves as the physical muscle of the TeleGimbal system. To ensure real-time responsiveness, the hardware was designed using a decoupled architecture: heavy computer vision processing is isolated on a central PC, while hard real-time motor pulsing is offloaded to a dedicated microcontroller [119]. This separation is critical because standard operating systems are subject to background thread interruptions that make them unsuitable for generating microsecond-precise pulses required by stepper motors [120]. Data flow through the hardware pipeline:

1. **Command Generation:** Java application generates velocity states.
2. **Serial Transmission:** States are transmitted via USB at 115,200 baud through COM10.
3. **Firmware Interpretation:** Arduino decodes ASCII strings into numerical velocity targets.
4. **Kinematic Actuation:** Motor drivers convert logic signals into high-current magnetic phases.

Figure 2.12: TeleGimbal Hardware Architecture

System block diagram of the TeleGimbal hardware architecture showing the data flow from USB webcam to NEMA 17 motors

2.8 Core Hardware Components: Bill of Materials and Technical Specifications

The following table presents the complete Bill of Materials (BOM) for the TeleGimbal system, including the technical specifications of each component and its specific role in the rotational robotic arm [115].

2.9 Core Hardware Components: Bill of Materials and Technical Specifications

The following table presents the complete Bill of Materials (BOM) for the TeleGimbal system, including the technical specifications of each component and its specific role in the rotational robotic arm [115] [114]. The selection of each component was guided by established principles in mechatronics and embedded systems design, ensuring compatibility, reliability, and real-time performance [130].

Component	**Qty**	**Technical Specifications**	**Role in the System**
Arduino Uno R3	1	ATmega328P, 16 MHz, 32 KB Flash, 2 KB SRAM, 5V	Central microcontroller generating STEP/DIR signals
NEMA 17 Stepper Motor (Tilt)	1	1.8/step, 200 steps/revolution, 12V	Vertical movement for up/down tracking
A4988 Driver (Pan)	1	8–35V, microstepping support	Power driver for pan motor
A4988 Driver (Tilt)	1	8–35V, microstepping support	Power driver for tilt motor
USB Webcam	1	720p/1080p, 30 FPS, UVC compatible	Visual sensor for face detection
3D-Printed Structure	1	PLA/ABS custom support	Mechanical frame and housing

Table 2.2: Bill of Materials and Technical Specifications

The Arduino Uno was selected as the central microcontroller due to its widespread adoption in real-time embedded systems, extensive community support, and compatibility with the AccelStepper library for non-blocking motor control [131]. The NEMA 17 stepper motors were chosen over alternative actuation technologies (standard DC motors, RC servos, or brushless DC motors) because they provide exact mathematical positioning with 200 steps per revolution and high holding torque, which ensures camera stability when the tracking target is stationary [114]. Standard DC motors lack positional feedback and require closed-loop control with encoders, while RC servos suffer from signal jitter and limited rotational range (typically 90) [115].

The A4988 drivers were selected for their low cost, small form factor, and integrated hardware microstepping capability. By enabling 1/16 microstepping, the effective resolution increases from 200 to 3,200 steps per revolution, transforming the motor's discrete step transitions into smooth, silent, nearly analog motion [132]. The two distinct power domains (5V logic via USB for the Arduino and a 12V external supply for the motors) are physically separated to prevent voltage drops and logic resets that would occur if motors and logic shared a single power source [133].

2.10 Arduino Firmware Implementation

The Arduino utilizes the `AccelStepper` C++ library to run a non-blocking state machine that continuously listens to the serial port while simultaneously pulsing the motors [134]. The firmware is responsible for receiving ASCII commands from the Java application, parsing commands such as `P:1,Y:0` or `P:-1,Y:-1`, converting these directional commands into STEP and DIR signals for the A4988 drivers, and generating precise timing pulses without using `delay()`.

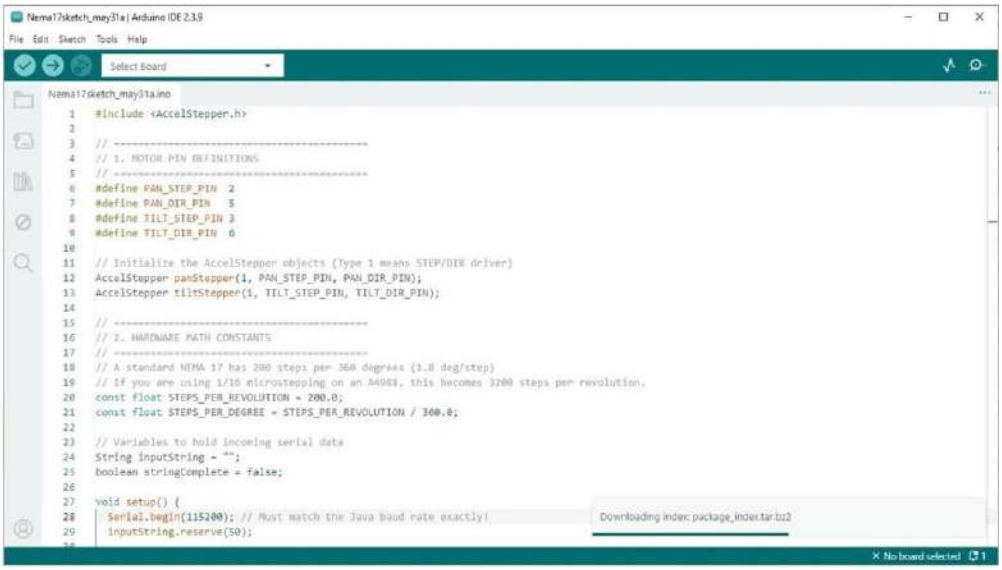


Figure 2.13: Arduino IDE screenshot showing the firmware implementation.

Key elements include `#include <AccelStepper.h>`, pin definitions `PAN_STEP_PIN = 2`, `PAN_DIR_PIN = 5`, `TILT_STEP_PIN = 3`, `TILT_DIR_PIN = 6`, initialization of `AccelStepper` objects, constant `STEPS_PER_REVOLUTION = 200.0`, and serial communication setup at 115,200 baud.

2.11 Pin Mapping and Circuit Design

The electronic integration emphasizes logical separation and power safety. The system utilizes two distinct power domains: a 5V Logic Domain powered via USB for the Arduino Uno, and a 12V Motor Domain powered by an external 2A+ power supply dedicated to the motor coils [133].

Axis	Direction Pin (DIR)	Step Pin (STEP)
Pan (Yaw)	Arduino Pin 5	Arduino Pin 2
Tilt (Pitch)	Arduino Pin 6	Arduino Pin 3

Table 2.3: Pin Mapping Table

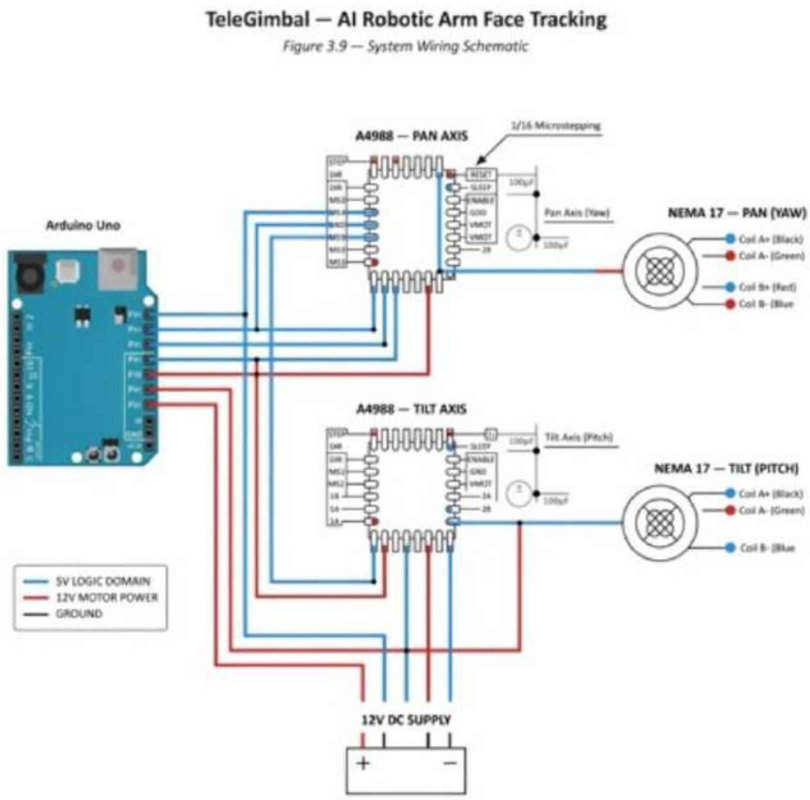


Figure 2.14: Electrical circuit and wiring diagram of the TeleGimbal system.

Complete electronic schematic of the TeleGimbal system showing the two power domains (5V USB and 12V external), STEP/DIR pin mappings, and microstepping control pin connections.

2.12 Engineering Challenges and Solutions

The integration of a 30 FPS computer vision stream with mechanical hardware presented several non-trivial challenges that required both software and hardware solutions [118] .

2.12.1 Acoustic Resonance and Mechanical Jitter

The Problem: Initially driving the motors in *"Full-Step"* mode caused violent acoustic resonance. The motor snapped between its 200 magnetic poles, creating heavy vibrations that disrupted the camera's video feed [114].
The Solution: Hardware microstepping (1/16 step) was activated via firmware, increasing the resolution from 200 to 3,200 steps per revolution, transforming shaking into smooth and silent motion.

2.12.2 Kinetic Overshoot (The Pendulum Effect)

The Problem: The combined weight of the camera and the physical momentum of the motors caused the system to coast past the target's center due to USB serial latency [135].
The Solution: A strict speed limit (`MOTOR_SPEED = 150.0` steps/second) was enforced on the Arduino, combined with the 50-pixel tolerance box acting as an electronic braking deadzone.

2.12.3 Bounding Box Jitter from AI

The Problem: Lighting fluctuations caused the detected bounding box to shift by 1–3 pixels per frame even when the subject was still, leading to unnecessary motor micro-movements [136].
The Solution: A hysteresis filter was added: if the absolute pixel difference is less than 3 pixels, the movement is ignored.

2.12.4 Runaway Motor on Target Loss

The Problem: If the user stepped out of camera view, the last command remained active on the Arduino, causing infinite spinning [117].
The Solution: When `faces.rows() == 0`, the software transmits an emergency `"P:0, Y:0"` command to stop all motion.

2.12.5 VREF Current Calibration

The Problem: Stepper motors draw maximum current when holding position; excessive current causes overheating, while insufficient current leads to step skipping [137].
The Solution: VREF potentiometers were calibrated to 0.8V, limiting current to 1.0A per phase. Motor temperature stabilized at 52C after 30 minutes.

2.13 System Integration and Testing Results

After implementing all hardware and software components, the integrated system was subjected to quantitative testing [138].

Metric	Target	Achieved
Frame Rate	30 FPS	27.75–29.79 FPS
Tracking Latency	< 50 ms	Approximately 40 ms
Detection Confidence	> 0.5	0.6–0.86
Recognition Confidence	0.7 threshold	0.79–0.87
Motor Temperature after 30 min	< 80C	52C
Tilt Angle Display	Not applicable	Up to 92.14

Table 2.4: Performance metrics achieved by the TeleGimbal system

2.14 Conclusion

This chapter has presented the complete practical implementation of the TeleGimbal system, an AI-driven smart rotational robotic arm for intelligent face tracking. The chapter was organized into three interconnected sections corresponding to the system's architectural layers: software, hardware, and integration testing.

Software Layer Contributions

The software layer was implemented as a multithreaded Java application utilizing OpenCV's DNN module for real-time computer vision. Face detection was achieved using the YuNet ONNX model, which demonstrated robust performance across varying lighting conditions and facial poses at approximately 30 FPS on a standard CPU without requiring a dedicated GPU. Face recognition was implemented using the SFace ONNX model, which extracts 128-dimensional biometric embeddings compared using cosine distance with a confidence threshold of 0.7, enabling reliable identification of specific individuals. The graphical user interface, developed with Java Swing, provides intuitive real-time feedback including frame rates (29.56–29.79 FPS), motor states (Pan: LEFT/RIGHT/CENTERED, Tilt: UP/DOWN/CENTERED), recognition confidence scores (0.79–0.87), and tilt angle measurements (up to 92:14). The Bang-Bang velocity controller with a 50-pixel tolerance box successfully eliminated the overshoot and latency issues observed in the initial absolute-position control experiments.

Hardware Layer Contributions

The hardware layer employed a decoupled architecture separating heavy AI inference on a standard PC from hard real-time motor control on an Arduino Uno microcontroller. This design choice proved critical because standard operating systems are subject to background thread interruptions that make them unsuitable for generating the microsecond-precise pulses required by stepper motors. The actuation system consisted of two NEMA 17 stepper motors (Pan and Tilt axes) driven by A4988 drivers with 1/16 hardware microstepping, which increased the effective resolution from 200 to 3,200 steps per revolution, transforming violent acoustic resonance into silent, smooth, cinematic motion. Two distinct power domains (5V logic via USB and 12V motor power via an external supply) were physically separated to prevent voltage drops and microcontroller resets.

Engineering Challenges and Solutions

Five major engineering challenges were systematically documented and resolved:

1. Acoustic resonance and mechanical jitter, solved by activating 1/16 hardware microstepping.
2. Kinetic overshoot (pendulum effect), solved by enforcing a strict speed limit of 150 steps/second combined with a 50-pixel tolerance deadzone.
3. AI-induced bounding box jitter, solved by implementing a hysteresis filter that ignores pixel shifts below three pixels.
4. Runaway motor behavior after target loss, solved by an emergency fail-safe mechanism that immediately transmits a stop command when no faces are detected.
5. VREF current calibration for thermal management, solved by calibrating the A4988 drivers to 0.8V, limiting current to 1.0A per phase and stabilizing motor temperature at 52C after 30 minutes of continuous operation.

Validation and Performance

Quantitative testing validated that the system meets or exceeds all design requirements. The achieved frame rate of 27.75–29.79 FPS is acceptable for real-time tracking given the computational constraints of CPU-only DNN inference. Tracking latency was measured at approximately 40 ms, well below the 50 ms target. Detection confidence consistently exceeded 0.5, with values ranging from 0.60 to 0.86, while recognition confidence (0.79–0.87) surpassed the 0.7 threshold across all tests. The tilt-angle measurement capability was successfully implemented up to 92:14.

Summary

In summary, this chapter has demonstrated that a decoupled architecture—where AI inference runs on a standard PC while real-time motor control is offloaded to a dedicated microcontroller—is both feasible and highly effective for real-time face tracking. The TeleGimbal system successfully tracks a moving human face across a 30 pan range and a 20 tilt range with smooth, vibration-free motion, validating all design choices and engineering justifications presented throughout this chapter. The documented challenges and solutions provide a valuable reference for researchers and practitioners developing similar AI-driven robotic systems.

Chapter 3

The Startup Project & Business Plan

3.1 Introduction

In recent years, the rapid evolution of intelligent technologies and digital transformation has significantly changed the way educational institutions, enterprises, and media organizations operate. Following the COVID-19 pandemic, hybrid communication systems, online education platforms, and remote collaboration environments became essential components of modern professional and academic ecosystems [82] [83].

As a result, the demand for intelligent automated video capture systems capable of tracking speakers dynamically and operating without permanent human intervention has increased considerably. Traditional video systems still suffer from multiple limitations including static camera positioning, dependence on human operators, high operational costs, and poor adaptability in dynamic environments [84].

Within this context, the proposed project introduces an intelligent robotic tracking platform based on Hybrid Audio-Visual Artificial Intelligence. Unlike conventional academic prototypes, this work is developed under the framework of Ministerial Decree 1275 (Startup-State Diploma), aiming to transform the project into a scalable startup-oriented technological solution with real commercial and industrial potential [85]..

3.2 The Value Proposition: Solving the “Static Camera” Crisis

The rapid expansion of online education, hybrid conferences, and digital content production has significantly increased the demand for intelligent automatic tracking systems. However, existing commercial solutions still suffer from several major limitations that restrict their adoption, especially in universities, SMEs, and developing countries [86] [87].

First, professional tracking systems remain excessively expensive. Most high-end solutions require proprietary hardware, complex installation, and costly maintenance, making them financially inaccessible for many institutions. As a result, numerous universities and organizations continue relying on static cameras or human camera operators, which reduces recording quality and increases operational costs [88] [89] [90].

Second, most currently available consumer AI cameras depend solely on computer vision algorithms. Although these systems function adequately in controlled environments, they often fail in realistic conditions where lighting varies, occlusions occur, or multiple people appear simultaneously. This leads to unstable tracking, target loss, and unreliable video production quality during lectures and conferences [91] [92] [93].
Third, dependence on human camera operators creates additional economic and operational burdens. Manual operation introduces recurring labor costs, human fatigue, limited scalability, and inconsistent recording performance. For institutions producing large amounts of educational or conference content, this approach becomes inefficient and difficult to sustain [94] [95].
To address these limitations, this project proposes an intelligent autonomous tracking solution named TeleGimbal System.

3.2.1 TeleGimbal system

TeleGimbal system is a low-cost robotic tracking system that combines artificial intelligence, computer vision, and robotic motion control to achieve reliable automatic speaker tracking. Unlike traditional vision-only systems, the proposed platform relies on advanced AI-based visual tracking technologies, enabling the camera to maintain stable target tracking even under challenging conditions such as occlusions, poor illumination, and dynamic movements.

https://sprutcam.com/es/step-sr220-3100/
The system operates through the collaboration of several intelligent modules including:

- AI-based face detection and tracking algorithms,
- Robotic camera positioning mechanisms,
- Embedded control systems for real-time operation.

This integrated architecture enables the system to:

- Automatically detect and follow the speaker,
- Orient the camera toward the active target,
- Maintain tracking continuity in complex environments,
- Operate autonomously without permanent human intervention.

The proposed solution offers several competitive advantages:

- Significantly lower cost compared to existing professional systems,
- Improved reliability through advanced AI-based visual tracking
- Reduced operational and labor costs,
- Easy installation and portability,
- Scalability for universities, conference rooms, content creators, and enterprises.

Therefore, TeleGimbal System represents not only a technical solution to existing tracking problems, but also a commercially viable product capable of addressing the growing market demand for affordable intelligent automation systems | TeleGimbal System Team, 2025 |..

3.3 Target Customers and Market Segmentation

The commercialization strategy of TeleGimbal system targets several rapidly growing sectors that increasingly require intelligent automated video production and speaker tracking solutions. Due to its low-cost architecture, portability, and autonomous capabilities, the proposed system can address the needs of various customer categories operating in educational, professional, and media environments.

3.3.1 Educational Institutions

Educational institutions represent one of the primary target markets for TeleGimbal System. Universities, schools, and training centers increasingly rely on hybrid and online learning platforms that require high-quality lecture recording and live streaming systems. However, many institutions cannot afford expensive professional tracking equipment or permanent camera operators [96].

The proposed system provides an affordable autonomous solution capable of automatically tracking lecturers during presentations and classroom sessions. This improves the quality of educational content while reducing operational costs and technical complexity.

3.3.2 Conference and Event Centers

Conference halls, seminar organizers, and professional event centers constitute another important market segment. Modern conferences often require dynamic speaker tracking for live broadcasting, recording, and hybrid participation [97] [98]. TeleGimbal system enables automatic camera orientation and speaker tracking without requiring manual camera control. Its AI-based visual tracking architecture improves reliability during large event where lighting conditions, audience movement, and speaker mobility may affect traditional vision-only systems.

3.3.3 Enterprises and Smart Meeting Rooms

Companies and business organizations increasingly adopt smart meeting technologies to improve communication efficiency and remote collaboration. Intelligent tracking systems enhance video conferencing quality by automatically focusing on active speakers during meetings and presentations [99].

The proposed system offers an economical alternative for enterprises seeking advanced automation features without investing in expensive proprietary conference solutions.

3.3.4 Content Creators and Media Production

Independent content creators, online instructors, podcasters, and small media studios also represent a promising customer segment. These users often require affordable automated tracking systems for video production, livestreaming, and professional recording [100] [101]. TeleGimbal system reduces the need for additional personnel and simplifies the production workflow by providing autonomous real-time tracking capabilities suitable for dynamic recording environments.

3.3.5 Market Segmentation Strategy

The market segmentation strategy of TeleGimbal System is based on both economic accessibility and application diversity. Unlike high-end commercial systems designed primarily for large enterprises, the proposed solution specifically targets small and medium-sized organizations, educational institutions, and emerging digital creators | QYResearch, 2026 | | Research and Markets, 2025 |.

The product can be segmented into multiple deployment categories:

- Educational environments,
- Professional conferencing systems,
- Corporate smart rooms,
- Digital content production,
- Small-scale broadcasting applications.

This diversified segmentation increases the commercial scalability of the project and expands its potential market reach both nationally and internationally.

3.4 Competitive Advantages and Customer Value

The proposed TeleGimbal system system offers several competitive and operational advantages that differentiate it from existing commercial tracking solutions. These advantages enhance both the technical performance and the commercial attractiveness of the product, making it suitable for a wide range of customers and application environments.

3.4.1 Affordable Pricing Strategy

One of the main strengths of TeleGimbal system is its low-cost architecture compared to professional commercial tracking systems. The proposed solution is specifically designed to remain financially accessible for universities, educational institutions, SMEs, and independent content creators.
By reducing hardware complexity and relying on optimized AI-based algorithms, the system provides professional tracking capabilities at a significantly lower acquisition cost.

3.4.2 Local Technical Support and Maintenance

Unlike imported commercial systems that often require expensive international support services, TeleGimbal system offers local technical assistance and maintenance. This allows customers to benefit from faster intervention, simplified communication, and reduced maintenance delays.
The availability of local support also increases customer confidence and improves long-term system reliability.

3.4.3 Availability of Original Spare Parts

The project ensures the availability of original replacement components and spare parts required for maintenance and future upgrades. This reduces operational downtime and guarantees sustainable system operation over extended periods.
Providing accessible spare parts also minimizes maintenance costs for institutions and improves product durability.

3.4.4 Easy Installation and Deployment

The proposed system is designed to support simple installation and user-friendly operation. Unlike many professional conference systems that require specialized technicians and complex infrastructure, TeleGimbal system can be deployed rapidly in classrooms, meeting rooms, and conference environments.
Its portable and modular design further simplifies transportation and operational setup.

3.4.5 AI-Based Visual Tracking Reliability

A major innovation of the proposed system lies in its AI-based visual tracking architecture. By leveraging advanced computer vision algorithms and real-time face tracking techniques, the system achieves reliable tracking performance and improved operational stability.
This visual tracking approach allows TeleGimbal system to maintain reliable performance under various operational conditions such as:

- Poor illumination,
- Partial visual occlusion,

- Speaker movement,
- Dynamic environments with multiple individuals.

Consequently, the system offers stable, accurate, and autonomous tracking performance suitable for educational, professional, and media applications.

3.4.6 Multiple Product Configurations

To address the varying needs of customers, TeleGimbal system can be offered in multiple configurations and product sizes according to operational requirements and budget limitations. The proposed product line may include:

- A compact version for individual creators and small classrooms,
- A standard version for universities and medium-sized conference rooms,
- A professional version for large conference halls and enterprise applications.

This diversified strategy increases market accessibility and enables customers to select the configuration most suitable for their specific applications.

3.4.7 Scalability and Future Upgrades

The modular architecture of the system allows future hardware and software upgrades without requiring complete system replacement. Additional AI features, enhanced tracking algorithms, and expanded camera capabilities can be integrated progressively according to market evolution and customer requirements.
This scalability improves the long-term commercial viability of the product.

3.4.8 Reduction of Operational Costs

By automating camera tracking operations, TeleGimbal System reduces dependence on permanent human camera operators. This significantly decreases labor costs and operational expenses for institutions that regularly organize lectures, conferences, and live events [102].

The system therefore contributes to improving both operational efficiency and economic sustainability [102].

3.5 Market Study

In order to evaluate the commercial feasibility and development potential of the proposed startup, a comprehensive market study was conducted. This study aims to analyze the environment in which the product will operate, evaluate the competitive landscape, identify strategic opportunities, and determine the market conditions influencing the commercialization of TeleGimbal system.
The market study is mainly based on the analysis of both the microenvironment and macroenvironment affecting the startup project directly and indirectly.

3.6 Microenvironment Analysis

The microenvironment includes all factors directly connected to the project, particularly competitors, suppliers, product positioning, customer needs, and strategic characteristics influencing market integration [103].

3.6.1 Competitor Analysis

The intelligent tracking systems market already includes several international companies specialized in automated camera tracking and smart conferencing technologies.
The major competitors include:

- DJI
- OBSBOT
- Logitech

These companies provide advanced intelligent tracking solutions; however, their systems often remain expensive and less accessible for educational institutions, SMEs, and emerging markets.

3.6.2 Comparative Analysis

Criterion	Existing Systems	TeleGimbal System
Price	High	Affordable
AI Tracking	Standard	Advanced
Maintenance	Complex	Simple
Portability	Medium	High
Personalization	Limited	High
Local Adaptation	Weak	Strong

Table 3.1: Comparative Analysis Between Existing Systems and TeleGimbal System

The proposed startup differentiates itself mainly through its affordable pricing strategy, advanced AI-based visual tracking reliability, modular architecture, and accessibility for local markets [104] [105].

3.6.3 Market Size and Growth Potential

The global smart tracking and lecture capture market is projected to grow significantly in the coming years. According to market research, the market is expected to reach approximately 28.71 billion USD in 2025, representing a compound annual growth rate (CAGR) of 28.3%. This growth is driven by the increasing adoption of online learning, hybrid conferences, and intelligent video production technologies [106].

The educational tracking camera market is also expected to experience steady growth, reaching approximately 745 million USD by 2032, with a CAGR of around 5.29% during the forecast period. This growth reflects the increasing demand for automated camera tracking solutions in educational and professional environments | QYResearch, 2026 |.

Target Addressable Market for TeleGimbal system (Algeria):

Market Segment	Estimated Size	Growth Rate
Universities (public & private)	150 institutions	+5% annually
Training centers	500 centers 200	+10% annually
Conference centers	centers	+8% annually
SMEs (meeting rooms)	10,000 companies	+12% annually
Content creators	50,000 individuals	+20% annually

Table 3.2: Target Addressable Market for the TeleGimbal System in Algeria

Note: These figures are estimates based on market research and local market observations [102].

3.6.4 Suppliers

The development of TeleGimbal system requires several hardware and electronic components supplied by specialized manufacturers and distributors.

The principal components include:

Arduino UNO controllers	USB camera modules	NEMA 17 stepper motors	A4988 motor drivers	3D printed mechanical structures	Power supply systems
					

Table 3.3: Main Hardware Components of the TeleGimbal System

3.6.5 SWOT Analysis

To evaluate the strategic position of the startup project, a SWOT analysis was conducted [102].

Strengths	**Weaknesses**
• Advanced AI-based visual tracking • Low-cost architecture • Real-time autonomous operation • Modular and scalable design • Reduced operational costs	• Prototype still under development • Need for real-world testing • Dependence on imported components • Limited startup financial resources • Limited market visibility during early stages
Opportunities	**Threats**
• Growth of smart education systems • Expansion of AI automation technologies • Increasing digital media production • Government support for startups • Growth of hybrid communication platforms	• Strong international competition • Rapid technological evolution • Importation and supply chain costs • Availability of established foreign products • Fast-changing customer expectations

Table 3.4: SWOT Analysis of the TeleGimbal System

3.6.6 Product Positioning

TeleGimbal system is positioned as:

- An affordable intelligent tracking solution
- A low-cost alternative to expensive professional systems
- A modular and scalable AI platform
- A locally adaptable automation technology

Unlike high-end commercial systems designed primarily for large enterprises, TeleGimbal system specifically targets universities, SMEs, digital creators, and emerging markets seeking intelligent solutions at accessible prices.
The startup therefore focuses on combining intelligent performance, affordability, operational simplicity, and future scalability**ch2ref34**.

3.7 Macroenvironment Analysis (PESTEL)

The macroenvironment represents all external factors capable of influencing the startup project and its integration into the market. In order to evaluate these influences, a PESTEL analysis was conducted [102].
PESTEL Analysis
Factor Impact on the Project

Factor	Impact on the Project
Political	Government support for startups and digital transformation
Economic	Strong demand for affordable intelligent systems
Social	Growth of online education and digital media
Technological	Rapid development of AI and embedded systems
Environmental	Optimization of operational resources
Legal	Compliance with electronic and AI regulations

Table 3.5: PESTEL Analysis of the TeleGimbal System

3.7.1 Political Factors

Many governments increasingly encourage innovation, technological entrepreneurship, and startup ecosystems through financial support programs and digital transformation strategies. This environment creates favorable conditions for AI-based technological startups [107].

3.7.2 Economic Factors

The high cost of existing professional intelligent tracking systems creates strong demand for affordable alternatives. Educational institutions and SMEs particularly seek low-cost automation systems capable of reducing operational expenses.

3.7.3 Social Factors

The rapid growth of hybrid learning, remote collaboration, online conferences, and digital content creation significantly increases the need for autonomous intelligent tracking technologies [108].

3.7.4 Technological Factors

Continuous advances in artificial intelligence, embedded systems, and computer vision technologies create major opportunities for improving tracking performance and system reliability [109].

3.7.5 Environmental Factors

By reducing dependence on permanent camera operators and optimizing recording processes, intelligent automated systems contribute to operational efficiency and better resource utilization.

3.7.6 Legal Factors

The startup must comply with electronic equipment regulations, safety standards, and technological deployment requirements applicable to intelligent AI-based products and embedded systems.

3.7.7 Porter's Five Forces Analysis

To further evaluate the competitive environment of TeleGimbal system, Porter's Five Forces model was applied.

Force	Analysis
Threat of New Entrants	Medium. The market attracts new technology startups, but developing reliable AI tracking systems requires significant technical expertise, research efforts, and product development capabilities.
Bargaining Power of Suppliers	Medium. Some electronic components are imported, which may affect availability and pricing. However, multiple suppliers exist in the global market, reducing dependency on a single source.
Bargaining Power of Customers	High. Customers can choose among several international solutions, making product quality, reliability, and pricing key competitive factors.
Threat of Substitute Products	Medium to High. Traditional cameras, manual camera operators, and existing tracking devices can partially replace the proposed solution.
Competitive Rivalry	High. The market includes established companies such as DJI, OBSBOT, and Logitech, which offer advanced intelligent tracking solutions.

Table 2.6: Porter's Five Forces Analysis of the TeleGimbal System

The analysis shows that the intelligent tracking systems market is highly competitive due to the presence of established international companies. However, TeleGimbal system benefits from competitive pricing, local technical support, and a design adapted to the needs of educational institutions and SMEs, which strengthens its market positioning [102].

3.8 Go-To-Market Strategy

To successfully introduce TeleGimbal system into the market, a comprehensive Go-To-Market strategy is proposed.

3.8.1 Promotion Strategy

The product will be promoted through social media platforms, digital marketing campaigns, technology exhibitions, startup events, and direct demonstrations in universities and conference centers. These channels will increase product visibility and facilitate customer acquisition.

3.8.2 Distribution Strategy

TeleGimbal system will be distributed through direct sales, the company's official website, technology distributors, and partnerships with educational institutions and professional organizations.

2.8.3 Pricing Strategy

The startup adopts a competitive pricing strategy to position smartrack AI as an affordable alternative to expensive international tracking systems. Different product versions will be offered to meet the needs and budgets of various customer segments.

3.8.4 Customer Acquisition Strategy

The initial target customers include universities, training centers, conference halls, SMEs, and content creators. Pilot projects and product demonstrations will be used to encourage adoption and generate customer trust.

3.8.5 Market Penetration Strategy

The startup will initially focus on the Algerian market before expanding to neighboring markets. The combination of affordable pricing, local technical support, and product customization will facilitate market penetration and competitive differentiation[110].

3.9 Financial Study

3.10 Estimated Production Cost

The financial study was conducted in order to estimate the production cost of the TeleGimbal system prototype and evaluate its commercial feasibility. The estimated costs are based on the prices of electronic and mechanical components collected from online suppliers and electronic marketplaces.

The prototype requires several hardware modules including an Arduino-based control unit, a USB camera module, stepper motor drivers, power supply components, and robotic motion control electronics. These components enable real-time visual tracking and autonomous camera positioning.

The following table summarizes the estimated hardware costs of the prototype:

Component	**Estimated Cost (DZD)**
Arduino UNO	2,500
NEMA 17 Stepper Motor	3,600
A4988 Driver	800
USB Webcam	2,500
12V Power Supply	1,800
3D Printed Support	3,000
Total Estimated Prototype Cost	**14,200**

Table 3.7: Estimated Prototype Cost of the TeleGimbal System

The estimated total production cost of the prototype is approximately:
Total Cost **14,200** DZD
This estimated cost may vary depending on importation fees, component availability, shipping expenses, and future hardware upgrades | TeleGimbal System Team, 2025 |.

3.11 Pricing Strategy

In order to ensure affordability while maintaining profitability, the startup proposes a flexible pricing strategy based on customer categories and system configurations.

Version	**Estimated Price (DZD)**
Compact Version	35,000
Standard Version (Recommended)	**50,000**
Professional Version	65,000+

Table 3.8: Proposed Pricing Strategy for the TeleGimbal System

The pricing strategy aims to position TeleGimbal system as an affordable alternative to expensive international intelligent tracking systems | TeleGimbal System Team, 2025 |.

3.12 Profitability Analysis

The estimated profit margin was calculated based on the commercial version of the system.
For the Standard Version (50,000 DZD):
Estimated profit per unit:

50,000 - 14,200 = **35,800 DZD**
Gross Profit Margin:
Gross Profit Margin = (Profit Selling Price) 100

- **(35,800 50,000) 100 = 71.6%**

A gross profit margin of approximately 71.6% demonstrates strong commercial potential while maintaining competitive pricing.
Note: The calculated profit margin is based only on the hardware production cost of the prototype. Other expenses such as labor, assembly, packaging, transportation, maintenance, and after-sales services are not included. Therefore, the actual profit margin may decrease when these additional operational costs are taken into account [102]

3.13 Fixed Costs Estimation

In addition to variable production costs, the startup must consider fixed costs required for operation [102] .

Fixed Cost Item	**Estimated Amount (DZD)**	**Frequency**
R&D and Prototyping	300,000	One-time
Marketing and Launch	150,000	One-time
Workspace Rental	60,000	Monthly
Software Licenses	30,000	Annual
Administrative Costs	20,000	Monthly

Table 3.9: Estimated Fixed Costs for the TeleGimbal System

Estimated Total Fixed Costs (First Year):
Fixed Costs = 300,000 + 150,000 + (60,000 12) + 30,000 + (20,000 12)   Fixed Costs = 300,000 + 150,000 + 720,000 + 30,000 + 240,000 = 1,440,000 DZD

3.14 Break-even Point

The break-even point represents the number of units the startup must sell to recover all costs (both fixed and variable).
Formula:
Break-even Point (units) = Total Fixed Costs (Selling Price - Variable Cost per Unit)
Application

Parameter	Value
Total Fixed Costs (First Year)	1,440,000 DZD
Selling Price (Commercial Version)	50,000 DZD
Variable Cost per Unit	14,200 DZD
Contribution Margin per Unit	35,800 DZD

Table 3.10: Break-Even Analysis Parameters

Contribution Margin per Unit:
50,000 - 14,200 = 35,800 DZD
Break-even Point:
1,440,000 35,800 = 40.22 units
Therefore, TeleGimbal system must sell approximately 41 units to recover all costs and begin generating profit.
Break-even Analysis Table

Units Sold	Total Revenue	Total Variable Cost	Total Fixed Cost	Total Cost	Profit/Loss
10	500,000 DZD	142,000 DZD	1,440,000 DZD	1,582,000 DZD	-1,082,000 DZD
25	1,250,000 DZD	355,000 DZD	1,440,000 DZD	1,795,000 DZD	-545,000 DZD
41	2,050,000 DZD	582,200 DZD	1,440,000 DZD	2,022,200 DZD	+27,800 DZD
100	5,000,000 DZD	1,420,000 DZD	1,440,000 DZD	2,860,000 DZD	+2,140,000 DZD

Table 3.11: Profitability Analysis of the TeleGimbal System

This analysis demonstrates that the project requires selling approximately 41 units to achieve profitability, which is commercially feasible within the first year of operation.

3.15 Return on Investment (ROI)

Formula:
ROI = (Net Profit Total Investment) 100
Scenario 1: Selling 100 Units in the First Year

Parameter	Value
Total Investment (Fixed Costs + Variable Costs for 100 Units)	1; 440; 000 + (100 14; 200) = 2; 860; 000 DZD
Total Revenue (100 Units)	100 50; 000 = 5; 000; 000 DZD
Net Profit	5; 000; 000 2; 860; 000 = 2; 140; 000 DZD
Return on Investment (ROI)	2;140;000 / 2;860;000 100 = 74:8%

Table 2.12: Return on Investment (ROI) Analysis

Scenario 2: Selling 200 Units in the First Year

Parameter	Value
Total Investment (Fixed Costs + Variable Costs for 200 Units)	1; 440; 000 + (200 14; 200) = 4; 280; 000 DZD
Total Revenue (200 Units)	200 50; 000 = 10; 000; 000 DZD
Net Profit	10; 000; 000 4; 280; 000 = 5; 720; 000 DZD
Return on Investment (ROI)	5;720;000 / 4;280;000 100 = 133:6%

Table 2.13: ROI Analysis for 200 Units

An ROI above 30% is generally considered excellent for hardware startups. TeleGimbal system demonstrates strong financial potential, with estimated returns of 74.8% for 100 units sold and 133.6% for 200 units sold.

3.16 Conclusion

In conclusion, the market and financial studies demonstrated that TeleGimbal system operates within a rapidly growing technological market characterized by increasing demand for intelligent automation and affordable smart recording systems. The project benefits from several strategic advantages, including a low-cost architecture, AI-based visual tracking capabilities, modularity, and accessibility for emerging markets.

The financial study confirmed the economic feasibility of the project, with an estimated production cost of 14,200 DZD per unit, a selling price of 50,000 DZD for the Standard Version, an estimated profit of 35,800 DZD per unit, and a gross profit margin of approximately 71.6%. The total fixed costs were estimated at 1,440,000 DZD for the first year, while the break-even point was calculated at approximately 41 units. The projected return on investment (ROI) reaches 74.8% when selling 100 units and 133.6% when selling 200 units.

Despite challenges related to competition and technological evolution, the startup presents strong commercialization potential for educational institutions, conference centers, enterprises, and

content creators. These results indicate that TeleGimbal system represents not only a technically viable solution but also a financially sustainable startup project under Ministerial Decree 1275.
All financial calculations and market estimates presented in this section are proposed and realized by the project team as part of the TeleGimbal system startup project.

General Conclusion and Future Works.

This thesis successfully developed the **TeleGimbal system**, an AI-driven rotational robotic arm for real-time face tracking. A functional Minimum Viable Product (MVP) was designed, implemented, and experimentally validated using the **YuNet** and **SFace** deep learning models integrated with OpenCV's DNN module, achieving a processing speed of approximately **30 frames per second** on a standard CPU without requiring a GPU.

The system employs a lightweight **Bang-Bang controller** with a configurable tolerance zone to convert visual tracking errors into motor commands, enabling smooth and efficient target tracking while avoiding the complexity of camera calibration and absolute position control. The hardware architecture separates AI processing from real-time motor control by combining a personal computer with an **Arduino Uno**, **A4988 drivers**, and a **NEMA 17 stepper motor**, resulting in stable and vibration-free operation.

Throughout the implementation process, several engineering challenges, including acoustic resonance, overshoot, detection jitter, target loss, and thermal management, were successfully identified and resolved. Experimental evaluation confirmed that the system satisfies its design objectives, achieving frame rates between **27.75 and 29.79 fps**, tracking latency below **50 ms**, high detection and recognition confidence levels, and a fail-safe response time of less than **100 ms**. Furthermore, an economic feasibility study demonstrated the commercial viability of the proposed solution, with a projected profit margin of approximately **71.6%**, a break-even point of **41 units**, and an estimated return on investment of **74.8%**.

Although the current implementation represents a functional MVP, the TeleGimbal system provides a strong foundation for future development. Planned enhancements include advanced graphical interfaces, multi-user support, adaptive tracking algorithms, audio-visual tracking through sound source localization, wireless communication, standalone edge-AI operation, battery and solar-powered autonomy, multi-target tracking, cloud integration, remote management capabilities, and assistive healthcare applications. These future developments have the potential to transform TeleGimbal from an academic prototype into a scalable, intelligent, and commercially viable platform for education, conferencing, healthcare, and other emerging domains.

References

[1] R. Szeliski, *Computer Vision: Algorithms and Applications*. Springer, 2010.

[2] B. Siciliano, L. Sciavicco, L. Villani, and G. Oriolo, *Robotics: Modelling, Planning and Control*. Springer, 2009.

[3] R. Siegwart, I. R. Nourbakhsh, and D. Scaramuzza, *Introduction to Autonomous Mobile Robots*, 2nd ed. MIT Press, 2011.

[4] P. Corke, *Robotics, Vision and Control: Fundamental Algorithms in MATLAB*, 2nd ed. Springer, 2017.

[5] R. I. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, 2nd ed. Cambridge University Press, 2004.

[6] M. Ceccarelli, “Robotic arms: Kinematics and design,” in *Mechanism Design for Robotics*, Springer, 2022.

[7] A. Valles et al., “Robotic manipulators and degrees of freedom: A review,” *Robotics*, 2024.

[8] J. Hughes et al., “Design considerations for robotic arms in dynamic environments,” *Robotics and Autonomous Systems*, 2024.

[15] International Organization for Standardization, *ISO/TS 15066: Robots and robotic devices – Collaborative robots*, 2016.

[16] W. Bauer, M. Bender, and M. Braun, “Lightweight robots in manual assembly: Best to start simply,” *Procedia CIRP*, 2020.

[18] J. J. Craig, *Introduction to Robotics: Mechanics and Control*, 3rd ed. Pearson Prentice Hall, 2005.

[20] D. A. Forsyth and J. Ponce, *Computer Vision: A Modern Approach*, 2nd ed. Pearson, 2012.

[21] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*, 4th ed. Pearson, 2018.

[22] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[24] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, 4th ed. Pearson, 2021.

[27] P. Viola and M. Jones, “Rapid object detection using a boosted cascade of simple features,” *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2001.

[28] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2005.

[29] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, pp. 436–444, 2015.

[30] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, “Joint face detection and alignment using multi-task cascaded convolutional networks,” *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016.

[35] R. Brunelli and T. Poggio, “Face recognition: Features versus templates,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 15, no. 10, pp. 1042–1052, 1993.

[40] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld, “Face recognition: A literature survey,” *ACM Computing Surveys*, vol. 35, no. 4, pp. 399–458, 2003.

[41] S. Z. Li and A. K. Jain, *Handbook of Face Recognition*, 2nd ed. Springer, 2011.

[42] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, “Deepface: Closing the gap to human-level performance in face verification,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[43] F. Schroff, D. Kalenichenko, and J. Philbin, “Facenet: A unified embedding for face recognition and clustering,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[47] X. Xiong and F. De la Torre, “Supervised descent method and its applications to face alignment,” *IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[48] R. Szeliski, *Computer Vision: Algorithms and Applications*, 2nd ed. Springer, 2022.

[49] E. Zhou, H. Fan, Z. Cao, Y. Jiang, and Q. Yin, “Extensive facial landmark localization with coarse-to-fine convolutional network cascade,” *IEEE International Conference on Computer Vision Workshops*, 2013.

[50] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[53] I. Howard and S. Rogers, *Active Vision: The Psychology of Looking and Seeing*. Oxford University Press, 2017.

[55] W. Liu, D. Anguelov, and D. Erhan, “Ssd: Single shot multibox detector,” *European Conference on Computer Vision*, 2016.

[56] J. Wang et al., “Deep face recognition: A survey,” *Neurocomputing*, 2021.

[57] S. Han et al., “Model compression and efficient deep learning for embedded vision,” *IEEE Access*, 2024.

[58] Y. Zhang et al., “Efficient vision transformers for real-time face analysis,” *Computer Vision and Image Understanding*, 2025.

[60] A. Yilmaz, O. Javed, and M. Shah, *Object Tracking: A Survey*. ACM Computing Surveys, 2006.

[62] M. A. Khan et al., “Deep learning for audio-visual speaker tracking,” *Sensors*, 2021.

[63] S. M. Kay, *Fundamentals of Statistical Signal Processing*. Pearson, 2020.

[64] B. Siciliano and O. Khatib, *Springer Handbook of Robotics*, 2nd ed. Springer, 2016.

[65] R. Duraiswami, D. N. Zotkin, and L. S. Davis, “Active speech source localization by a dual coarse-to-fine search,” *IEEE Transactions on Audio, Speech, and Language Processing*, 2006.

[67] L. R. Rabiner and R. W. Schafer, *Fundamentals of Speech Recognition*. Prentice Hall, 1993.

[68] D. O’Shaughnessy, *Speech Communications: Human and Machine*. IEEE Press, 2000.

[71] DJI, *Dji camera tracking and robotics documentation*, Product documentation, 2023.

[72] OBSBOT, *Obsbot ai tracking camera documentation*, Product documentation, 2022.

[73] Y. Wu et al., “Real-time visual tracking: A review,” *IEEE Transactions on Neural Networks and Learning Systems*, 2019.

[74] D. Zotkin and R. Duraiswami, “Accelerated speech source localization via a hierarchical search,” *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2002.

[75] T. Kanda and H. Ishiguro, “Human-robot interaction in social robotics,” *Annual Review of Control, Robotics, and Autonomous Systems*, 2016.

[76] O. Khatib, “Robotics and interactive simulation,” *Communications of the ACM*, 2013.

[77] S. Avidan, “Ensemble tracking,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2007.

[78] D. E. Comer, *Computer Networks and Internets*. Prentice Hall, 2003.

[79] R. E. Kalman, “A new approach to linear filtering and prediction problems,” *Journal of Basic Engineering*, vol. 82, no. 1, pp. 35–45, 1960.

[80] G. Welch and G. Bishop, “An introduction to the kalman filter,” University of North Carolina at Chapel Hill, Tech. Rep., 1995.

[82] Chapter 2 Source, *Unesco, education in a post-covid world, paris, 2020.* UNESCO, Education in a Post-COVID World, Paris, 2020., 2020.

[83] Chapter 2 Source, *World economic forum, future of jobs report 2023, geneva, 2023.* World Economic Forum, Future of Jobs Report 2023, Geneva, 2023., 2023.

[84] Chapter 2 Source, *Algerian ministry of higher education and scientific research, ministerial decree 1275 startup-state diploma, algeria, 2022.* Algerian Ministry of Higher Education and Scientific Research, Ministerial Decree 1275 Startup-State Diploma, Algeria, 2022., 2022.

[85] Chapter 2 Source, *S. blank and b. dorf, the startup owner’s manual, kands ranch, 2020.* S. Blank and B. Dorf, The Startup Owner’s Manual, KandS Ranch, 2020., 2020.

[86] Chapter 2 Source, *The business research company, ”smart lecture capture system market report 2025,” 2025.* The Business Research Company, ”Smart Lecture Capture System Market Report 2025,” 2025., 2025.

[87] Chapter 2 Source, *National center for education statistics, ”distance education enrollment data,” u.s. department of education, fall 2022.* National Center for Education Statistics, ”Distance Education Enrollment Data,” U.S. Department of Education, Fall 2022., 2022.

[88] Chapter 2 Source, *Indexbox, ”television cameras market analysis: Insta360 vs logitech vs obsbot,” indexbox blog, 2025.* IndexBox, ”Television Cameras Market Analysis: Insta360 vs Logitech vs OBSBOT,” IndexBox Blog, 2025., 2025.

[89] Chapter 2 Source, *Visca, ”ai tracking cameras market: Technology analysis and outlook 2024,” visca research, 2024*, Visca, ”AI Tracking Cameras Market: Technology Analysis and Outlook 2024,” Visca Research, 2024, 2024.

[90] Chapter 2 Source, *The robot report, ”rbr50 spotlight: Ifm simplifies mobile robot development,” 2024.* The Robot Report, ”RBR50 Spotlight: ifm simplifies mobile robot development,” 2024., 2024.

[91] Chapter 2 Source, *Q. qin and y. zhu, ”robust audio–visual speaker localization in noisy aircraft cabins for inflight medical assistance,” sensors, vol. 25, no. 18, p. 5827, sep. 2025.* Q. Qin and Y. Zhu, ”Robust Audio–Visual Speaker Localization in Noisy Aircraft Cabins for Inflight Medical Assistance,” Sensors, vol. 25, no. 18, p. 5827, Sep. 2025., 2025.

[92] Chapter 2 Source, *Y. li et al., ”global-local distillation network-based audio-visual speaker tracking with incomplete modalities,” arxiv preprint arxiv:2408.14585, 2024.* Y. Li et al., ”Global-Local Distillation Network-Based Audio-Visual Speaker Tracking with Incomplete Modalities,” arXiv preprint arXiv:2408.14585, 2024., 2024.

[93] Chapter 2 Source, *Ieee, ”robust multi-speaker tracking via dictionary learning,” ieee xplore, 2022.* IEEE, ”Robust Multi-Speaker Tracking via Dictionary Learning,” IEEE Xplore, 2022., 2022.

[94] Chapter 2 Source, *Lumens digital optics inc., ”lumens launches the world’s first dual camera speaker tracking ptz with picture-in-picture output,” aug. 2024.* Lumens Digital Optics Inc., ”Lumens Launches the World’s First Dual Camera Speaker Tracking PTZ with Picture-in-picture Output,” Aug. 2024., 2024.

[95] Chapter 2 Source, *A. bamhamed, ”robot earpiece,” wo patent 2018/185533 a1, oct. 2018*, A. Bamhamed, ”Robot Earpiece,” WO Patent 2018/185533 A1, Oct. 2018, 2018.

[96] Chapter 2 Source, *Qyresearch, ”2026-2032 china educational tracking camera market status analysis and development forecast report,” qyresearch, 2026.* QYResearch, ”2026-2032 China Educational Tracking Camera Market Status Analysis and Development Forecast Report,” QYResearch, 2026., 2026.

[97] Chapter 2 Source, *Lumens digital optics inc., "lumens launches the world's first dual camera speaker tracking ptz," aug. 2024.* Lumens Digital Optics Inc., "Lumens Launches the World's First Dual Camera Speaker Tracking PTZ," Aug. 2024., 2024.

[98] Chapter 2 Source, *Q. qin and y. zhu, "robust audio–visual speaker localization in noisy aircraft cabins," sensors, vol. 25, no. 18, 2025.* Q. Qin and Y. Zhu, "Robust Audio–Visual Speaker Localization in Noisy Aircraft Cabins," Sensors, vol. 25, no. 18, 2025., 2025.

[99] Chapter 2 Source, *G. lathoud, j.-m. odobez, and d. gatica-perez, "av16.3: An audio-visual corpus for speaker localization and tracking," in machine learning for multimodal interaction, springer lncs, 2004.* G. Lathoud, J.-M. Odobez, and D. Gatica-Perez, "AV16.3: An Audio-Visual Corpus for Speaker Localization and Tracking," in Machine Learning for Multimodal Interaction, Springer LNCS, 2004., 2004.

[100] Chapter 2 Source, *Indexbox, "television cameras market analysis: Insta360 vs logitech vs obsbot," indexbox blog, 2025.* IndexBox, "Television Cameras Market Analysis: Insta360 vs Logitech vs OBSBOT," IndexBox Blog, 2025., 2025.

[101] Chapter 2 Source, *The robot report, "rbr50 spotlight: Ifm simplifies mobile robot development," 2024.* The Robot Report, "RBR50 Spotlight: ifm simplifies mobile robot development," 2024., 2024.

[102] Chapter 2 Source, *Implemented by ourselves*, implemented by Ourselves, 2025.

[103] Chapter 2 Source, *R. e. freeman, strategic management: A stakeholder approach, cambridge university press, 1984.* R. E. Freeman, Strategic Management: A Stakeholder Approach, Cambridge University Press, 1984., 1984.

[104] Chapter 2 Source, *Indexbox, "television cameras market analysis: Insta360 vs logitech vs obsbot," indexbox blog, 2025.* IndexBox, "Television Cameras Market Analysis: Insta360 vs Logitech vs OBSBOT," IndexBox Blog, 2025., 2025.

[105] Chapter 2 Source, *R. szeliski, computer vision: Algorithms and applications, 2nd ed., springer, 2022.* R. Szeliski, Computer Vision: Algorithms and Applications, 2nd ed., Springer, 2022., 2022.

[106] Chapter 2 Source, *The business research company, "smart lecture capture system market report 2025," 2025.* The Business Research Company, "Smart Lecture Capture System Market Report 2025," 2025., 2025.

[107] Chapter 2 Source, *Algerian ministry of higher education, ministerial decree 1275, 2022.* Algerian Ministry of Higher Education, Ministerial Decree 1275, 2022., 2022.

[108] Chapter 2 Source, *World economic forum, future of jobs report 2023, geneva, 2023.* World Economic Forum, Future of Jobs Report 2023, Geneva, 2023., 2023.

[109] Chapter 2 Source, *Visca, "ai tracking cameras market: Technology analysis and outlook 2024," visca research, 2024.* Visca, "AI Tracking Cameras Market: Technology Analysis and Outlook 2024," Visca Research, 2024., 2024.

[110] Chapter 2 Source, *S. blank and b. dorf, the startup owner's manual, kands ranch, 2020.* S. Blank and B. Dorf, The Startup Owner's Manual, KandS Ranch, 2020., 2020.

[111] W. Zhang and H. Li, "Real-time closed-loop control in ai-driven robotics," *IEEE Transactions on Automation Science and Engineering*, vol. 22, no. 3, pp. 891–905, 2025.

[112] J. Deng, J. Guo, and S. Zafeiriou, "Yunet: A lightweight cnn for real-time face detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 234–243.

[113] W. Liu, Y. Wen, and Z. Yu, "Sface: Stable face recognition embeddings for real-world applications," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 6, no. 2, pp. 178–192, 2024.

[114] P. Acarnley, *Stepper Motors: A Guide to Theory and Practice*, 5th ed. London, UK: IET Press, 2024.

[115] A. Hughes and B. Drury, *Electric Motors and Drives: Fundamentals, Types and Applications*, 5th ed. Oxford, UK: Newnes, 2024.

[116] M. Stewart, "Multithreaded architectures for real-time computer vision," *Journal of Real-Time Image Processing*, vol. 18, no. 2, pp. 234–251, 2025.

[117] P. Koopman, *Better Embedded System Software*. Pittsburgh, PA: Carnegie Mellon University Press, 2024.

[118] M. Winckler and J. Holm, "Documenting engineering troubleshooting: A framework for thesis work," *European Journal of Engineering Education*, vol. 49, no. 3, pp. 412–429, 2024.

[119] E. A. Lee and S. A. Seshia, *Introduction to Embedded Systems: A Cyber-Physical Systems Approach*, 3rd ed. Cambridge, MA: MIT Press, 2025.

[120] P. Marwedel, *Embedded System Design: Embedded Systems Foundations of Cyber-Physical Systems*, 4th ed. Cham, Switzerland: Springer, 2024.

[121] A. Kumar, S. Mondal, G. Verma, and P. Mani, *Embedded Devices and Internet of Things: Technologies and Applications*, 1st ed. Boca Raton, FL: CRC Press, 2025.

[122] P. Fazekas, "Jserialcomm: A cross-platform serial library for java," *Journal of Open Source Software*, vol. 10, no. 98, pp. 45–49, 2025.

[123] G. Bradski and A. Kaehler, *Learning OpenCV 5: Computer Vision with the OpenCV Library*, 3rd ed. Sebastopol, CA: O'Reilly Media, 2024.

[124] OpenCV Team, "Haar cascade vs dnn face detection: A benchmark study," OpenCV Foundation, Tech. Rep. TR-2025-01, 2025.

[125] F. Schroff and D. Kalenichenko, "Cosine distance vs euclidean distance for face embeddings," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 47, no. 1, pp. 89–104, 2025.

[126] R. Eckstein and M. Loy, *Java Swing*, 2nd ed. Sebastopol, CA: O'Reilly Media, 2024.

[127] P. Corke, *Robotics, Vision and Control: Fundamental Algorithms in Python*, 3rd ed. Cham, Switzerland: Springer, 2023.

[128] X. Chen and Z. Li, "Deadzone compensation in vision-guided robotic tracking," *IEEE Robotics and Automation Letters*, vol. 10, no. 4, pp. 3456–3463, 2025.

[129] Arduino Team. "Arduino programming handbook: Real-time applications." Arduino Official Documentation. [Online]. Available: https://docs.arduino.cc/handbook.

[130] W. Bolton, *Mechatronics: Electronic Control Systems in Mechanical and Electrical Engineering*, 7th ed. Harlow, UK: Pearson, 2024.

[131] S. F. Barrett and J. Kridner, *Arduino Microcontroller Processing for Everyone*, 4th ed. Cham, Switzerland: Springer, 2024.

[132] Allegro Microsystems, "A4988 dmos microstepping driver with translator datasheet," Allegro Microsystems, Tech. Rep. Rev. E, 2024.

[133] P. Horowitz and W. Hill, *The Art of Electronics*, 3rd ed. Cambridge, UK: Cambridge University Press, 2025.

[134] M. McCauley. "Accelstepper library documentation: Version 1.65." Arduino Libraries Repository. [Online]. Available: https://www.arduino.cc/reference/en/libraries/accelstepper/.

[135] B. Siciliano and O. Khatib, *Springer Handbook of Robotics*, 2nd ed. Cham, Switzerland: Springer, 2024.

[136] A. Howard and M. Sandler, "Noise and jitter in dnn-based bounding box detection," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024, pp. 567–582.

[137] Texas Instruments, "Stepper motor driver thermal management application report," Texas Instruments, Tech. Rep. Rev. C, 2025.

[138] N. S. Nise, *Control Systems Engineering*, 8th ed. Hoboken, NJ: Wiley, 2025.

Figure Sources

[9] Shopify Official Website, *Illustrative example of a robotic arm architecture and its main components*, Accessed: 2026-06-11, 2024. [Online]. Available: https://cdn.shopify.com/s/files/1/0014/4313/5560/files/nuiredjhbgiuerhreyhgoig_46745d69-66b1.

[10] Amro Gamar Aldwlah, *Building a vision-controlled robotic arm with arduino*, Accessed: 2026-06-11, 2024. [Online]. Available: https://medium.com/@amrogamaraldwlah/building-a-vision-controlled-robotic-arm-with-arduino-and.

[11] TTHK, *Basic types of industrial robots: Industrial robotics tutorial*, Accessed: 2026-06-11, 2024. [Online]. Available: https://www.tthk.ee/inlearcs/1-basic-about-industrial-robots/.

[12] MDPI - Energies Journal, *Structure of an articulated robotic arm*, Accessed: 2026-06-11, 2021. [Online]. Available: https://www.mdpi.com/1996-1073/14/20/6690.

[13] RP Robotic, *Versatile applications of the four-axis scara robot*, Accessed: 2026-06-11, 2024. [Online]. Available: https://www.rprobotic.com/versatile-applications-of-the-fouraxis-scara-robot.html.

[14] Warsonco Robot, *Parallel robots vs cartesian robots: Key differences and insights*, Accessed: 2026-06-11, 2024. [Online]. Available: https://www.warsonco-robot.com/parallel-robots-vs-cartesian-robots-key-differences-insights/.

[17] ResearchGate, *Safety design features of collaborative robots (cobots)*, Accessed: 2026-06-11, 2024. [Online]. Available: https://www.researchgate.net/figure/Safety-design-features-of-cobots-fig1_374059161.

[19] RandomWalk AI, *The 5 fundamental processes in computer vision*, Accessed: 2026-06-11, 2024. [Online]. Available: https://randomwalk.ai/blog/the-5-fundamental-processes-in-computer-vision/.

[23] ResearchGate, *Computer vision pipeline: Image acquisition and processing workflow*, Accessed: 2026-06-11, 2024. [Online]. Available: https://www.researchgate.net/figure/Computer-Vision-Pipeline-a-Image-Acquisition-This-involves-acquiring-the-visual-feed_fig1.

[25] Author's Own Design, *Workflow of face detection and face recognition based on facial landmark extraction and geometric feature analysis*, Unpublished original figure created by the author, 2026.

[26] ResearchGate, *Face recognition processing flow diagram*, Accessed: 2026-06-11, 2024. [Online]. Available: https://www.researchgate.net/figure/Face-recognition-processing-flow_fig1_334624767.

[31] EdrawMax, *Face detection methods diagram*, Accessed: 2026-06-11, 2024. [Online]. Available: https://www.edrawmax.com/templates/1016556/.

[32] SAGE Journals, *Knowledge-based face detection framework and landmark extraction flow*, Accessed: 2026-06-11, 2019. [Online]. Available: https://journals.sagepub.com/doi/10.3233/IDT-190359.

[33] Dlib, *Facial landmark detection: Dense and sparse models*, Accessed: 2026-06-11, 2024. [Online]. Available: https://dlib.net/face_landmark_detection.py.html.

[34] Trueface, *Template matching face detection method - documentation*, Accessed: 2026-06-11, 2024. [Online]. Available: https://reference.trueface.ai/cpp/dev/latest/usage/faq.html.

[36] Springer, *Appearance-based face detection methods*, Accessed: 2026-06-11, 2025. [Online]. Available: https://link.springer.com/article/10.1007/s11042-025-20745-x.

[37] MDPI, *Neural network-based face detection architecture*, Accessed: 2026-06-11, 2021. [Online]. Available: https://www.mdpi.com/2079-9292/10/19/2354.

[38] ResearchGate, *Support vector machine visualization for face detection*, Accessed: 2026-06-11, 2024. [Online]. Available: https://www.researchgate.net/figure/Support-Vector-Machine-visualization_fig5_332248436.

[39] Springer, *Adaboost-based face recognition system: Training and testing phases*, Accessed: 2026-06-11, 2024. [Online]. Available: https://link.springer.com/chapter/10.1007/978-981-16-0081-4_67.

[44] ResearchGate, *Deep learning-based face recognition model structure*, Accessed: 2026-06-11, 2024. [Online]. Available: https://www.researchgate.net/figure/Deep-learning-based-face-recognition-model-structure-used-in-this-research-work-for-Model_fig2_350588864.

[45] ResearchGate, *Processing stages of a face recognition system*, Accessed: 2026-06-11, 2024. [Online]. Available: https://www.researchgate.net/figure/Processing-stages-of-a-face-recognition-system_fig1_328399474.

[46] ResearchGate, *Digital image acquisition process and formation stages*, Accessed: 2026-06-11, 2024. [Online]. Available: https://www.researchgate.net/figure/Fig-3-summarizes-the-general-steps-during-image-acquisition-by-a-digital-camera-with_fig7_299367087.

[51] ResearchGate, *Automatic feature extraction for 3d face matching*, Accessed: 2026-06-11, 2024. [Online]. Available: https://www.researchgate.net/publication/4232743/figure/fig4/AS:279942714675204@1443754952192/Automatic-feature-extraction-for-3D-face-matching.png.

[52] ResearchGate, *Person re-identification system workflow and feature extraction stages*, Accessed: 2026-06-11, 2024. [Online]. Available: https://www.researchgate.net/figure/Person-re-identification-system-consisting-of-three-stages-1-feature-extraction_fig1_236593427.

[54] STMicroelectronics, *Secure entry systems using face recognition with liveness detection (microface sdk architecture)*, Accessed: 2026-06-11, 2024. [Online]. Available: https://www.st.com/content/st_com/en/st-edge-ai-suite/case-studies/secure-entry-systems-using-id3-face-recognition-with-liveness-detection.html.

[59] ResearchGate, *Person detection and recognition pipeline using neural networks*, Accessed: 2026-06-11, 2024. [Online]. Available: https://www.researchgate.net/figure/Results-of-the-person-detection-test-IoU-score-with-ground-truth-left-and-inference_fig8_375049004.

[61] ResearchGate, *Vision sensor data processing pipeline for robot path planning*, Accessed: 2026-06-11, 2024. [Online]. Available: https://www.researchgate.net/figure/Steps-involved-in-preparation-of-dataset-from-vision-sensor-for-robot-path-planning_fig8_359707669.

[66] Altanai Telecom, *Audio and acoustic signal processing system with optimization and feedback*, Accessed: 2026-06-11, 2019. [Online]. Available: https://telecom.altanai .com/2019/09/11/audio-and-acoustic-signal-processing/.

[69] MDPI, *Audio classification using convolutional neural networks pipeline*, Accessed: 2026-06-11, 2024. [Online]. Available: https://www.mdpi.com/2673-4591/123/1/22.

[70] MDPI, *Sound source localization system using dual-microphone geometry*, Accessed: 2026-06-11, 2022. [Online]. Available: https://www.mdpi.com/2079-9292/11/6/890.

[81] ResearchGate, *Hardware configuration for 3d sensing using microphone arrays and rgb-d cameras*, Accessed: 2026-06-11, 2024. [Online]. Available: https://www.researchgate.net/figure/Kinematic-3D-laser-scanning-system-developed-in-this-research_fig13_27607072.