2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

Integrating Artificial Vision and Wearable Robotics: Adaptive Assistance enabled by Manipulation Context Awareness

Sandro Ferrari[1], Emanuele Aimi[1], Francesco Missiroli[1], Federico Masiero[1], Maura Casadio[2], and Lorenzo Masia[1*]

***Abstract*— Occupational exoskeletons are emerging as a promising solution for industrial applications, providing support to reduce fatigue and the risk of musculoskeletal disorders. One of the main challenges limiting their widespread adoption is that most existing devices cannot deliver real-time, adaptable, and context-aware assistance. This paper presents the first fully vision-driven control strategy for a bimanual upper-limb soft exoskeleton, enabling adaptive assistance during industrial tool manipulation. The approach integrates three modules: tool recognition and segmentation, hand tracking with gesture recognition, and a fusion layer that ensures reliable understanding of the manipulation context. This allows modulation of lifting assistance in real time according to the weight of the grasped object. Experiments with human participants demonstrated that the proposed approach reduces biceps activation by more than 50% compared to the no-support condition, while operating in real time on embedded hardware. The method is robust to hand–object occlusions, camera repositioning, and dynamic environments, demonstrating its practicality for industrial deployment. Overall, this work establishes vision-based control as a scalable solution for ergonomic, adaptive exoskeletons that enhance safety and productivity in demanding workplaces.**

***Index Terms*— Computer Vision, Human-Robot Interaction, Wearable Robotics, Adaptive Assistance Systems, Object Segmentation, Weight Estimation.**

I. Introduction

Exoskeletons for industrial applications were originally conceived to augment human strength through bulky and heavy devices designed to provide their operator with the ability to lift extremely heavy loads [1], [2]. More recently, the primary goal of exoskeletons shifted from endowing the operator with "superhuman strength," to reduce muscle fatigue and musculoskeletal disorders arising from repetitive and physically demanding tasks [3], [4]. Rigid and bulky exoskeletons have shown clear limitations in terms of usability, comfort, and integration into occupational settings [4], [5], motivating the development of exosuits: soft, lightweight devices that provide more ergonomic and less intrusive support [6], [7]. Several research studies and arising commercial products have demonstrated that the use of exosuits reduces muscle effort, promoting ergonomics and mitigating the risk of musculoskeletal disorders [8], [9], [10], [11].

[1]S. Ferrari, E. Aimi, F. Missiroli, F. Masiero, and L. Masia are with the Munich Institute for Robotics and Machine Intelligence (MIRMI), Department of Computer Engineering, School of Computation, Information and Technology, Technical University of Munich (TUM), Munich, Germany,
[2]M. Casadio is with the DIBRIS, University of Genova, 16145 Genoa, Italy
[*]Corresponding author: Lorenzo Masia (lorenzo.masia@tum.de)

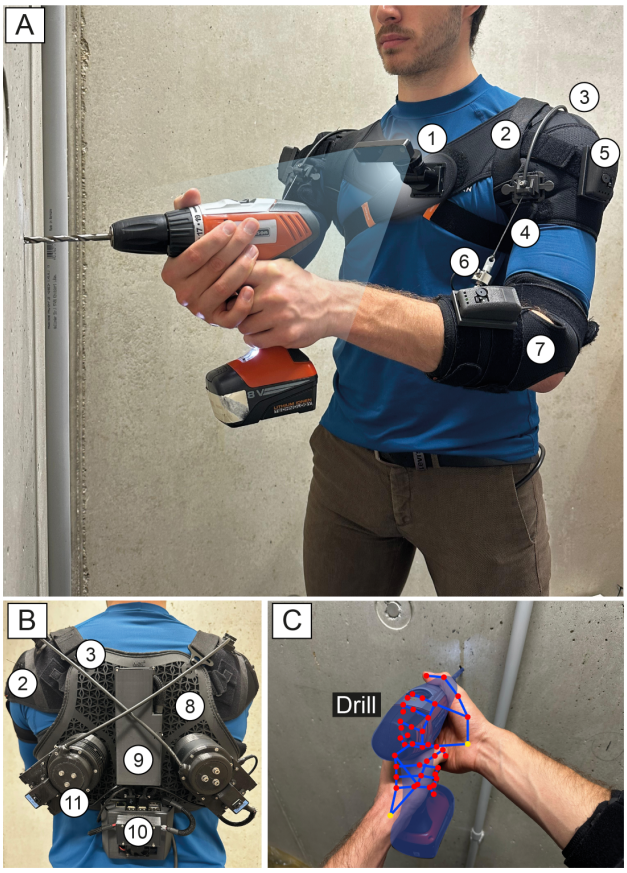

Fig. 1. System overview. A) Bimanual upper-limb soft exoskeleton with integrated camera vision for manipulation context awareness. A,B) The device comprises: 1) OAK D-Lite RGB-D camera, 2) shoulder orthosis, 3) Bowden cable, 4) artificial tendon, 5) IMU sensors, 6) load cell, 7) forearm orthosis, 8) back support, 9) battery, 10) NVIDIA Jetson Nano, 11) actuation module. C) Illustrative camera view of the exoskeleton with real-time detection of grasped objects.

Wearable robots can be either passive, providing constant support, or active, with the ability to modulate assistance according to both the operator's physical characteristics (e.g., weight and height) and motion intent. Current research focuses on advancing these technologies through adaptive, context-aware assistance [12]. In particular, the capability to modulate support based on the operator's interaction with the environment remains an open challenge in exoskeleton research, especially in industrial pipeline and warehouses [13], [14].

Assistance modulation can be tackled by including wearable sensing, such as electromyography (EMG) and force

myography (FMG) [11], [15], [16]. However, these methods are often impractical in industrial settings due to challenges in sensor placement, signal degradation over time, and the necessity of time-consuming calibrations [17], [13], [18], [19], [20]. An alternative direction, suitable for cooperative tasks, aims to integrate collaborative robotic arms and exoskeletons in the loop to dynamically adjust assistance and enforce ergonomics [21].

Computer vision has emerged as a powerful tool for modulating assistance in wearable robotics, as it enables contextual interpretation and a deeper understanding of the operator's interaction with the environment. Vision-based perception systems have successfully been employed to recognize terrain in lower-limb exoskeletons [22], [23] and were explored more recently to modulate assistance in lifting and manipulation tasks by integrating other sensor information (e.g., contact events from pressure sensors) [24]. Approaches relying only on vision are particularly attractive, as they avoid the need for additional skin- or hand-mounted sensors, thereby preserving simplicity of use and freedom of movement.

In this study, we introduce a novel control strategy that relies solely on vision to modulate lifting assistance in a bimanual soft upper-limb exoskeleton during industrial tool manipulation. The proposed pipeline comprises three main modules: tool recognition and segmentation, hand tracking with gesture recognition, and a fusion layer that integrates these outputs to interpret the manipulation context. A crucial feature of our system is its robustness to hand–object occlusions, ensuring reliable performance in realistic manipulation scenarios. Each tool category is associated with a nominal weight, which is used to modulate assistance in real time when an object of that category is grasped. Experiments demonstrate that this approach reduces operator muscle effort by more than 50% regardless of the grasped tool. In real industrial scenarios, this vision system is compatible with integration into a broad range of upper-limb exosuits, with the potential to further reduce fatigue and mitigate the risk of musculoskeletal disorders associated with prolonged tool usage. The main contributions of this work are threefold:

- **Vision-based adaptive control**: The first fully vision-driven system for real-time modulation of assistance in a bimanual industrial exoskeleton for the upper limbs, validated through experiments with human subjects.
- **Context-aware perception pipeline**: A modular vision pipeline that combines tool recognition and hand tracking with a fusion module designed to understand the manipulation context in realistic scenarios.
- **Occlusion-robust dataset**: A new dataset of annotated industrial tools, specifically designed to train segmentation networks robust to hand–object occlusions (dataset availability: `https://universe.roboflow.com/worktooldataset/worktooldataset-gqkjq`).

This study paves the way for future research on intelligent, context-aware adaptation in wearable robots, offering a practical and scalable solution to assist workers in repetitive and physically demanding industrial tasks.

II. Design

The bimanual soft exoskeleton employed in this study is a cable-driven system comprising two electromechanical actuation modules, each dedicated to supporting a single arm (Fig. 1A,B). Each motor actuates an artificial tendon attached to the user's forearm, with the aim of supporting the elbow joint.

The soft exoskeleton consists of a flexible back support and textile orthoses that wrap around the shoulders and forearms. The back support houses the battery (14.8 V, 3700 mAh), and two actuation modules, each consisting of a control unit (Arduino MKR 1010 WiFi, Arduino, Italy), and a motor (T-Motor AK60-6, 24 V, 6:1 planetary gear reduction, CubeMars, T-MOTOR, China).

Inertial sensors (IMU, Bosch BNO055, Gerlingen, Germany) are mounted on the shoulders and forearms to estimate relative arm kinematics and, in particular, to compute the elbow flexion angle. Simultaneously, a load cell (ZNLBM-1, 20 kg max, Bengbu Zhongnuo Sensor, China) measures the interaction torque between the tendon and the anchor point sewn in the corresponding forearm brace. Data streaming from the sensors in the textiles and the control unit is established via Bluetooth Low Energy (BLE) using dedicated microcontrollers (Feather nRF52 Bluefruit, Adafruit Industries, New York City, USA).

The soft exoskeleton also includes a vision system comprising a Jetson Nano (4 GB, NVIDIA Jetson Nano, Santa Clara, CA, USA) mounted on the back support and an RGB-D camera (OAK-D Lite, Luxonis, USA) mounted on the chest. Wired data transmission is used between Jetson Nano, camera and control units.

III. Control

Our control strategy is based on a gravity-compensation controller that operates independently for each arm and is augmented with a manipulation-aware control layer leveraging computer vision. This layer integrates information about grasped objects, particularly their estimated mass, to adapt the assistance provided by the exosuit.

The perception pipeline that informs this controller comprises three main modules: *Object Segmentation*, *Hand Tracking*, and *Manipulation Awareness*. The first two modules run in parallel to provide real-time information about objects within the camera's field of view and the user's hand configuration. Their outputs are fused by the *Manipulation Awareness* module, which infers the manipulation context and estimates the mass of the grasped object, whose mass is then used to modulate the exoskeleton's *Gravity Compensation Controller*, enabling context-aware assistance.

A. Object Segmentation

The role of this module is to map each image pixel to the semantic class of the object to which it belongs. For

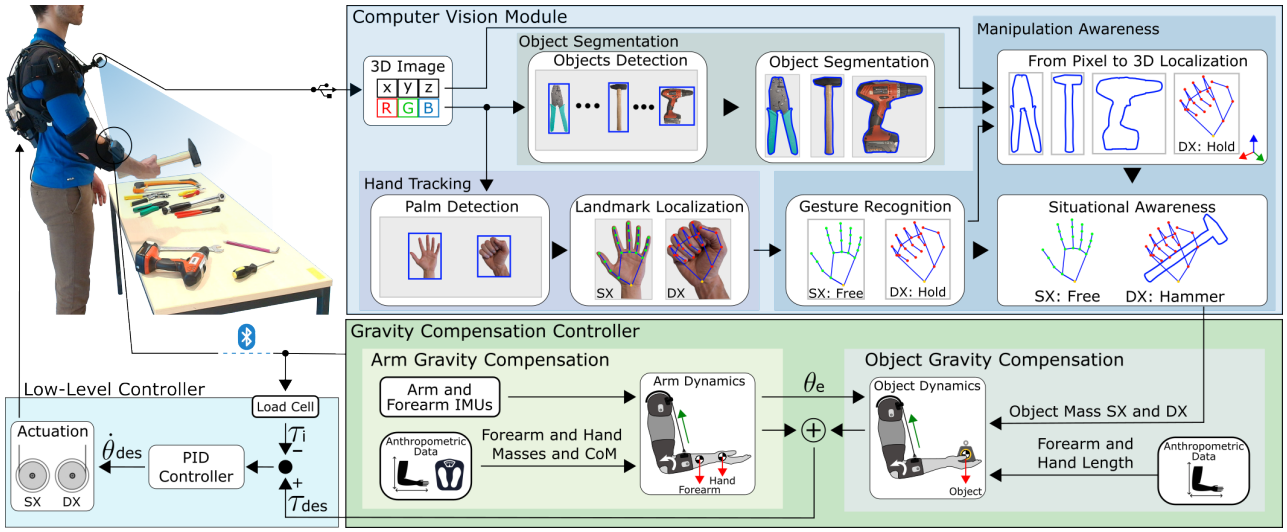

Fig. 2. Control scheme of the vision-based exoskeleton. The architecture includes a Computer Vision Module, a Gravity Compensation Controller, and a Low-Level Controller. RGB-D camera streams are processed to detect regions of interest (hand or object), enabling parallel hand tracking and object segmentation. By combining localized hand landmarks, segmented object regions, and 3D depth data, the Computer Vision Module infers manipulation awareness, distinguishing whether the hand is free or holding an object, and outputs the object's mass and the holding hand to the Gravity Compensation Controller. The latter, personalized to the user's anthropometry, uses object mass and arm kinematics (θ_e, $\dot{\theta}_e$, $\ddot{\theta}_e$) to compute the desired torque τ_{des}, to support the elbow joint. This reference torque is then fed into the Low-Level Controller, which employs a closed-loop PID controller to determine the motor speed $\dot{\theta}_{des}$ based on torque sensor feedback τ_i.

this purpose, we adopted a lightweight segmentation architecture derived from the YOLO family of models, known as `YOLOv8n-seg`. This model extends the standard object-detection framework by jointly predicting object bounding boxes, class probabilities, and per-pixel segmentation masks in a single inference step. The network employs a compact backbone for feature extraction, a path-aggregation neck to preserve multi-scale spatial information, and a segmentation head that combines anchor-free detection with dynamic mask generation. As a result, the architecture enables real-time instance-level segmentation while maintaining robustness to partial occlusions and variations in viewpoint. To maximize FPS on the embedded platform, we deployed the module on the Jetson Nano and ran inference on its onboard GPU, using batch size 1 and fixed input resolution.

In this study, eleven categories of hand tools commonly used in industrial contexts were selected as target objects: Crimper (0.39 kg), Drill (1.63 kg), Hammer (0.93 kg), Hexkey (0.14 kg), Nipper (0.34 kg), Pliers (0.20 kg), Punch (0.38 kg), Saw (0.54 kg), Screwdriver (0.09 kg), Spanner (0.29 kg), and Torque Wrench (0.89 kg).

To address hand–object occlusions, the network has been trained on a newly created pixel-level annotated dataset, including free and grasped objects.

B. Hand Tracking

This module is responsible for detecting the user's hands in real-time and estimating the configuration of their principal joints and left–right discrimination. We employed a lightweight model derived from the `MediaPipe Hands` framework, which implements a cascaded architecture optimized for low-latency applications. The method operates in two stages: (i) a palm detector identifies an oriented bounding box around the hand, leveraging anchor-free regression to ensure robustness under rotation and scale changes; (ii) within the cropped region, a landmark regression network predicts the spatial coordinates of 21 keypoints corresponding to finger joints and the palm structure. The combination of palm detection and a specialized landmark regressor substantially reduces computational overhead while maintaining subpixel accuracy in joint localization.

To maximize the overall throughput of the perception pipeline, this module was not executed on the central embedded platform but directly on the OAK-D Lite camera, whose integrated Myriad X VPU was used to run the optimized hand-tracking network in FP16 precision. Model weights were quantized and compiled into a device-specific blob format, and input dimensions were fixed to satisfy hardware constraints. This offloading strategy assigns hand tracking to the camera, freeing the Jetson Nano's GPU to run object segmentation and manipulation awareness at higher frequencies.

C. Manipulation Awareness

This module integrates information from both the *Object Segmentation* and *Hand Tracking* modules to determine whether the user is actively grasping an object. The segmentation module provides the semantic class of all objects in the camera's field of view together with their associated pixel masks, while the hand-tracking module outputs the pixel coordinates of hand landmarks. Based on the geometric configuration of these landmarks, the algorithm evaluates finger flexion to infer whether the hand is open or closed.

When the hand is detected as closed, the 2D coordinates of the hand landmarks and the recognized objects in the scene are projected into 3D space using the information provided

by the camera. A spatial proximity algorithm interprets their 3D positions, determines whether the hand is in contact with an object, and, if so, identifies the specific object being grasped. The mass associated with the detected object class is then transmitted from the computer vision module to the *Gravity Compensation Controller* independently for the right and left arms, allowing each arm to compensate for the weight of the object it is holding. If both hands are recognized as grasping the same object, the controller automatically divides the estimated mass equally between the two arms.

To increase the algorithm robustness against false detections of grasping and release, a windowing approach is used. Furthermore, if the user moves one hand out of the camera's field of view while holding an object, or if the camera view becomes temporarily occluded, the module retains the last valid grasp state in memory.

To achieve real-time performance, the *Hand Tracking* module is executed directly on the camera, while the *Object Segmentation* and *Manipulation Awareness* modules are parallelized on separate threads. Code optimization and parallelization strategies improved the performance of the entire perception pipeline from an average of approximately 3 fps to about 10 fps, both measured at an image resolution of 512×288. These results account for the full pipeline running in conjunction with the exoskeleton's control loop.

D. Gravity Compensation Controller

The contribution of the vision system extends the *gravity compensation* strategy implemented in the exoskeleton. Without vision, assistance is limited to compensating for the weight of the forearm, estimated through user anthropometry. IMU sensors placed on the arm and forearm provide kinematic measurements, while a load cell along the tendon measures the interaction torque, which is used to infer the user's intent to flex or extend the elbow.

With the integration of the vision module, the assistance model also accounts for the weight of grasped objects. By approximating arm dynamics with a rigid link formulation, the desired elbow torque is approximated using a simplified biomechanical model as:

$$\tau_{\text{des}} = (\mathrm{m}_f \, \mathrm{x}_{f,\text{CoM}} + \mathrm{m}_{\text{object}} \, \mathrm{x}_{h,\text{CoM}}) \, g \sin\theta_e + \ddot{\theta}_e \, \mathrm{x}_{f,\text{CoM}}^2 \, \mathrm{m}_f + \ddot{\theta}_e \, \mathrm{x}_{h,\text{CoM}}^2 \, \mathrm{m}_{\text{object}}, \quad (1)$$

where $\mathrm{m}_f = 0.022\,\mathrm{M_u}$ is the combined mass of the forearm and hand derived from body mass $\mathrm{M_u}$, $\mathrm{m}_{\text{object}}$ is the external object mass estimated by the vision system, $\mathrm{x}_{f,\text{CoM}} = 0.099\,\mathrm{H_u}$ is the distance from the elbow to the forearm–hand center of mass derived from height $\mathrm{H_u}$, and $\mathrm{x}_{h,\text{CoM}} = 0.201\,\mathrm{H_u}$ is the distance from the elbow to the hand/object center of mass. Here, g denotes gravitational acceleration and θ_e is the elbow angle measured by the IMUs.

The gravitational terms represent the torques due to the forearm–hand system and the object, while the inertial terms account for their respective moments of inertia; Coriolis effects are negligible and omitted. Finally, torque tracking is achieved with a PID-based admittance controller that takes τ_{des} as reference and closes the loop on the interaction torque τ_i, where τ_i is estimated from the cable-tension measurements provided by the load cell.

IV. Experiments

Two experimental protocols were carried out: a preliminary test to validate the vision system in isolation and an integration test to assess its effects when combined with the exoskeleton.

A. Vision Module Test

The vision system was evaluated for grasp detection and object classification with ten participants wearing a chest-mounted camera. Eleven objects were arranged on a table, and each subject performed ten grasps per object in random order. Trials began from a resting posture, followed by grasping, free arm movements, returning objects to the table, and returning to rest (Fig. 4a). The vision module operated online during the test, while synchronized video recordings were stored for offline analysis.

B. Integration Test

Nine healthy participants (mass: 76.0 kg, IQR 7.0; height: 1.80 m, IQR 0.06; age: 26.0 years, IQR 3.0) with no musculoskeletal or neurological disorders completed this test. The study followed the Declaration of Helsinki and was approved by the Ethics Committee of Heidelberg University (S-311/2020).

The task simulated manual manipulation of three objects (crimper, hammer, drill) of increasing weight. Two conditions were tested: *Without Exo* (baseline) and *Vision On* (assistance modulated by object type). Trials followed a standardized sequence of approach, grasp, lift, five-second static hold at 90° elbow flexion, lowering, and return, paced by auditory cues. Familiarization trials preceded testing to ensure timing consistency.

Each participant completed 30 trials (five repetitions per object per condition) in a single session. Conditions were randomized, and a 30-minute rest separated them to minimize fatigue.

System impact was evaluated through biceps brachii EMG, recorded following SENIAM guidelines. At session start, maximum voluntary contractions (MVCs) were collected for normalization. EMG and vision outputs were synchronously sampled at 1 kHz and stored for offline analysis.

V. Data Analysis

This section describes the analyses performed on the experimental data, organized into three parts: vision system evaluation, assessment of exoskeleton assistance, and statistical testing.

A. Computer Vision

The vision system was evaluated separately in a preliminary test and in the integration experiment.

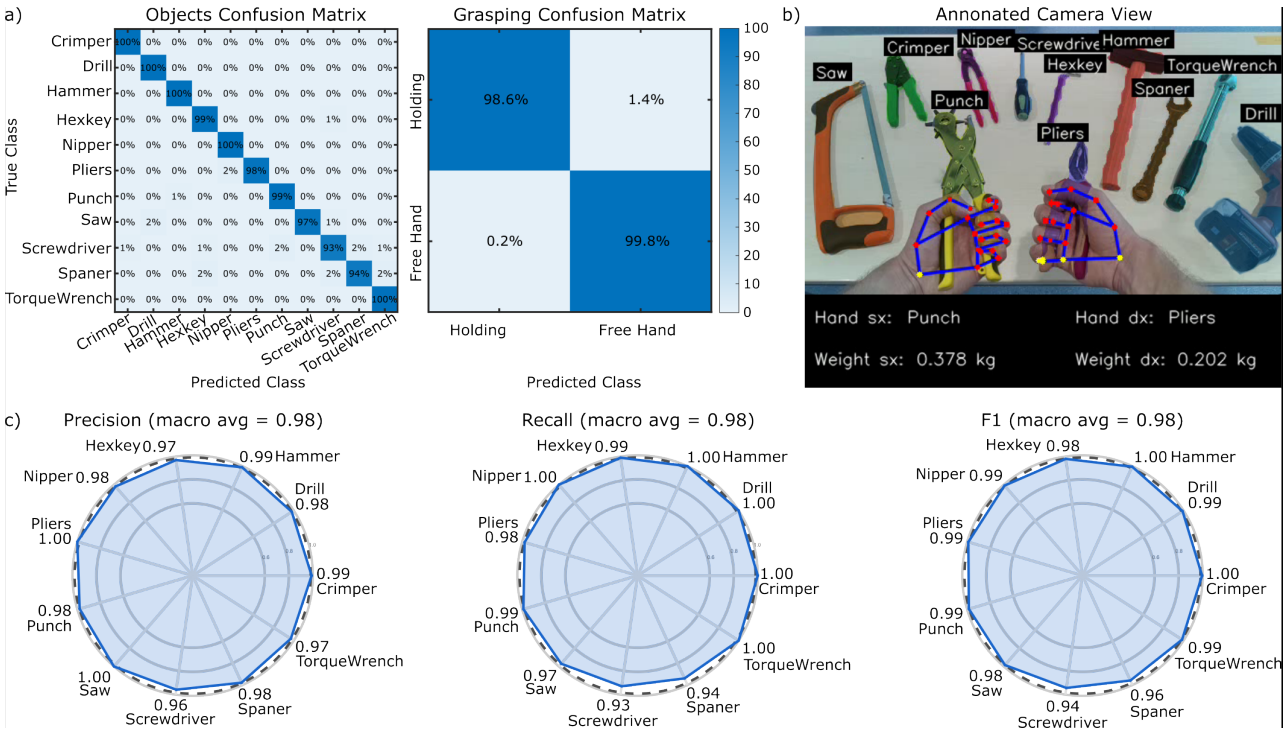

Fig. 3. Results of the preliminary vision-only experiment. A) Confusion matrices showing object classification and grasp detection accuracies. B) Example frame captured from a participant during the experiment. C) Spider plots illustrating, from left to right, the Precision, Recall and F1 score for object detection.

1) Preliminary Experiment: Each repetition (grasping an object) was treated as one evaluation of grasp detection and one of object recognition. For both tasks, we computed confusion matrices, precision, recall, and F1-scores. Object recognition was evaluated only when participants actively held an object, excluding background or table items. Ground truth labels were obtained by manual annotation of video recordings.

2) Integration Experiment: During the integration test, the system was evaluated at the frame level while participants manipulated objects with the exosuit, under challenging conditions such as partial occlusion or objects leaving the camera's field of view. Confusion matrices, precision, recall, and F1-scores were computed across all frames. At the trial level, we measured the percentage of repetitions with perfect grasp-state recognition; for trials with errors, we reported the mean percentage of correctly classified frames (±SEM). Ground truth was again determined via manual annotation. Additionally, the frame rate (FPS) of the integrated perception pipeline was measured to assess temporal efficiency, and classification accuracy was reported per participant to evaluate inter-subject consistency.

B. Adaptive Assistance

To assess the effect of assistance on muscle effort, EMG signals from the biceps brachii (Delsys Trigno system) were processed with a fourth-order Butterworth band-pass filter (15–450 Hz), rectified, and low-pass filtered at 6 Hz to obtain the linear envelope. Signals were normalized to each participant's MVC and quantified using the root mean square (RMS) across the entire manipulation (lifting, static holding, lowering). We therefore focused on biceps EMG, since the elbow-flexion task primarily recruits biceps and shoulder stabilizers, and the exoskeleton provides assistance exclusively at the elbow joint.

The biological torque at the elbow was computed as the difference between the total torque required for the movement (estimated by inverse dynamics of a second-order pendulum) and the assistive torque delivered by the exoskeleton (detailed in Section III-D and smoothed with a 0.1 s moving-average filter). From these time series, we extracted mean biological torque over the full manipulation.

1) Statistical Analysis: For each participant, normalized RMS EMG values and mean biological torque were averaged across five repetitions for each *Condition* (*Without Exo*, *Vision On*) × *Object* (crimper, hammer, drill) combination. Data normality was tested with the Shapiro–Wilk test ($\alpha = 0.05$). Since assumptions were violated, pairwise Wilcoxon signed-rank tests were applied. Statistical significance was set at $p < 0.05$, and effect sizes were reported alongside p-values.

VI. Results

A. Vision System Performance

1) Preliminary Experiment: Object and grasping accuracy proved always above 93% and 98.6%, respectively (Fig. 3a). At the repetition level, grasp detection achieved a precision, recall, and F1-score of 99.8%, 98.6%, and 99.2% for "Holding", and 98.6%, 99.8%, and 99.2% for "Free Hand". Object

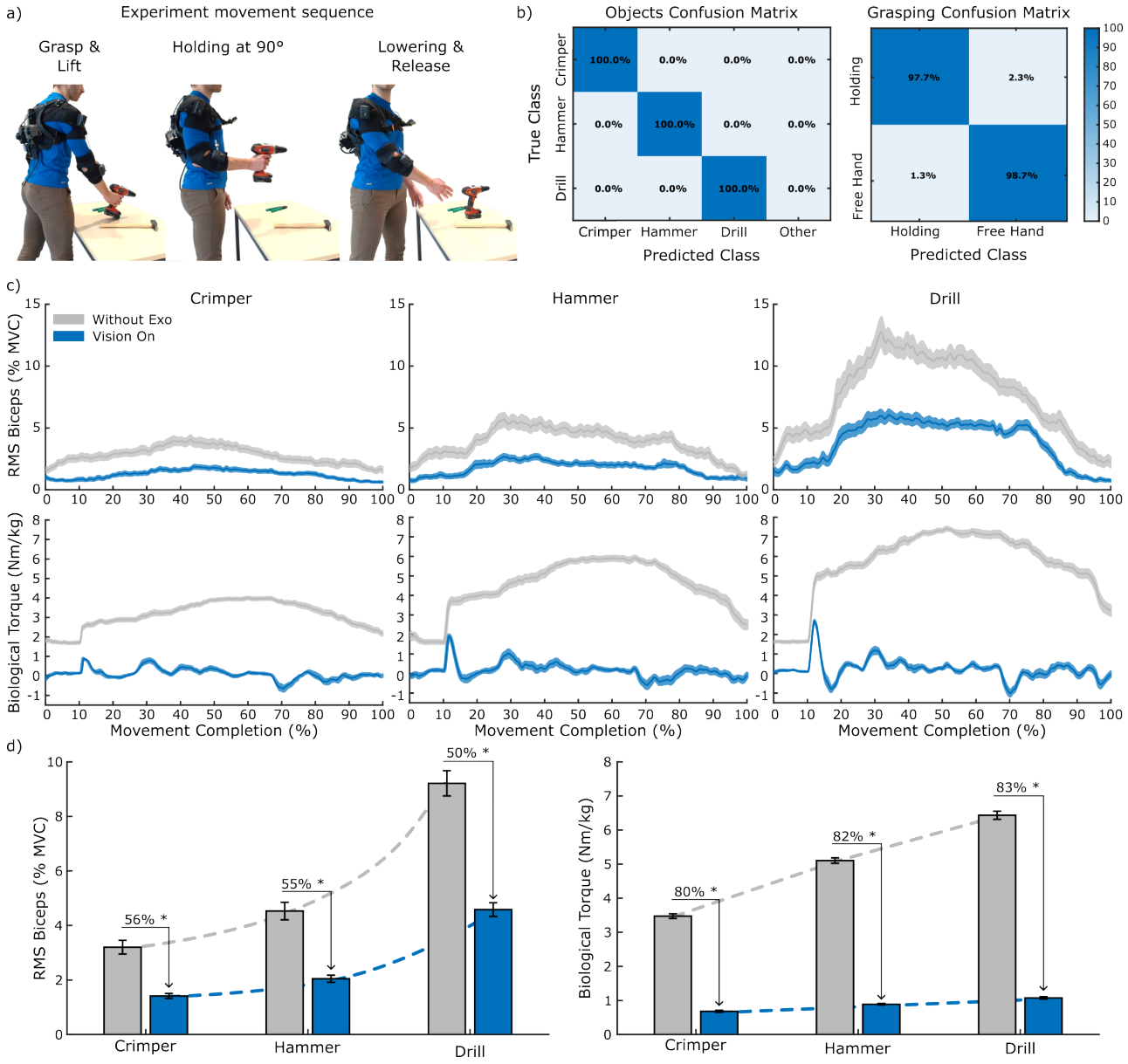

Fig. 4. Results of the Integration Experiment. A) Overview of the participant's motion sequence: grasping and lifting an object, holding it with the elbow at 90°, returning to the start position, and releasing it. B) Confusion matrices of object classification and grasp detection accuracies during the trials involving the soft exoskeleton. C) Inter-subject averaged RMS of biceps EMG activity, normalized to the participant's MVC, plotted against movement completion time for the three tested objects (left: crimper, middle: hammer, right: drill). Blue and gray curves represent mean values (± standard error) without exoskeleton and with vision on, respectively. D) Participants' RMS of biceps EMG activity (left) and biological torque (right) under assisted and unassisted conditions.

recognition yielded a precision of 98%, recall of 98%, and F1-score of 98% across the eleven tested categories (Fig. 3c).

2) Integration Experiment: At the frame level, grasp detection reached 98.7% precision, 97.7% recall, and 98.2% F1-score for "Holding",and 97.7% precision, 98.7% recall, and 98.2% F1-score for "Free Hand". The corresponding confusion matrix is reported in Fig. 4b. Object recognition during grasping phases reached 100% precision, 100% recall, and 100% F1-score across all three object classes, despite partial occlusion of the grasped object.

At the trial level, 90.37% of repetitions showed perfect grasp-state recognition across all frames. For repetitions with at least one error, the mean percentage of correctly classified frames was 84.23 ± 8.36% (mean ± SEM). The integrated perception pipeline operated at 10.44 ± 1.09 frames/s (mean ± std), confirming real-time feasibility.

B. Adaptive Assistance

1) Muscular Activation: Across participants, RMS activation decreased in the *Vision On* condition compared to the *Without Exo* condition (Fig. 4c). For the crimper, muscle activation was $3.20 \pm 0.25\%$ MVC without exosuit and $1.41 \pm 0.09\%$ MVC with vision on, corresponding to a relative reduction of 55.93%. For the hammer, values were

$4.53 \pm 0.32\%$ MVC and $2.04 \pm 0.13\%$ MVC (–54.96%). For the drill, activation decreased from $9.21 \pm 0.46\%$ MVC to $4.58 \pm 0.25\%$ MVC (–50.27%).

The Shapiro–Wilk test indicated that normality assumptions were not satisfied ($p = 0.0046$), so non-parametric statistical comparisons were applied. The Wilcoxon signed-rank test revealed a significant main effect of the delivery of the assistance regardless of the grasped object ($p = 0.0039$, $r_{rb} = -1.000$, $N = 9$), confirming lower muscle activation in the *Vision On* condition. Pairwise Wilcoxon comparisons confirmed significantly lower activation in the *Vision On* condition for each object: crimper ($p = 0.035$, raw $p = 0.0117$, $r_{rb} = -0.911$), hammer ($p = 0.0117$, raw $p = 0.0039$, $r_{rb} = -1.000$), and drill ($p = 0.0117$, raw $p = 0.0039$, $r_{rb} = -1.000$).

2) Biological Torque: Mean biological torque over the entire manipulation movement decreased in the *Vision On* condition compared to the *Without Exo* condition. For the crimper, torque dropped from 3.48 ± 0.07 N·m to 0.68 ± 0.03 N·m (–80.46%). For the hammer, it decreased from 5.10 ± 0.08 N·m to 0.89 ± 0.02 N·m (–82.55%). For the drill, torque was reduced from 6.43 ± 0.12 N·m to 1.08 ± 0.03 N·m (–83.20%).

The Shapiro–Wilk test indicated that normality assumptions were not satisfied ($p = 0.0010$), so non-parametric tests were again applied. The Wilcoxon signed-rank test revealed a significant main effect of *Condition* on mean biological torque ($p = 0.0039$, $r_{rb} = -1.000$, $N = 9$), confirming lower torque in the *Vision On* condition. Pairwise Wilcoxon comparisons confirmed significantly lower torque in the *Vision On* condition for each object: crimper, hammer, and drill ($p = 0.0117$, raw $p = 0.0039$, $r_{rb} = -1.000$, $N = 9$ for all).

VII. Discussion

Control for upper limb exoskeletons and prostheses is still an open challenge because of the high kinematics redundancy of the human arm and hand [25], which results in large motion variability depending on the task and the surrounding environment. Therefore, context awareness is crucial for enhancing the control of next-generation wearable robots. Our paper, together with other recent works in the wearable robotics field [24], [26], [27], supports that vision could be a game-changing solution for adaptive assistance modulation and shared-autonomous control. In this perspective, we present the integration of a real-time vision module into a soft bimanual upper limb exoskeleton for context-based adaptive assistance.

Our vision module showed excellent performance in both grasp detection and object classification, achieving high accuracy, precision, and recall. This robustness is particularly relevant given that it was obtained with an embedded, real-time setup, demonstrating its practical applicability in wearable systems. Notably, our approach did not involve the use of any extra sensor, i.e., it was based only on egocentric visual data. A further strength of the proposed vision system is its reliance on 3D scene data. This enables simultaneous localization of both hands and all objects not only in image coordinates but also in metric space, thereby allowing unambiguous identification of which hand has grasped which object. Purely pixel-based methods may misinterpret background objects as being closer than the grasped object, whereas the use of depth information makes the system robust to such projection artifacts. Moreover, because the system is based on segmentation rather than bounding-box detection, background elements are excluded, resulting in highly accurate 3D localization. To improve robustness, the system preserves the memory of the last grasped object even if the hand leaves the scene or an unexpected occlusion occurs. Another practical advantage is that both hand and object localization are carried out in the camera's relative reference frame, eliminating the need for calibration and stable camera placement.

Ultimately, the integration of the vision module into the soft exoskeleton led to a consistent and significant reduction of biceps activation during the manipulation of objects with different masses, with reductions exceeding 50% compared to the no-support condition. Including additional EMG recording sites would provide a more comprehensive characterization of muscle co-activation patterns; however, this was beyond the scope of the present study. Overall, these findings confirm that context aware information represents a solid alternative to physiological or skin-/hand-mounted sensors to deliver adaptive assistance.

The proposed vision modulation approach is particularly promising for future industrial applications. By relying solely on egocentric 3D scene data, the system enables precise grasp detection and object classification without the need for additional obtrusive sensors or calibration procedures. This makes it inherently scalable and easy to integrate into dynamic workplace environments where flexibility and reliability are critical. Its segmentation-based strategy further enhances robustness by excluding background clutter and maintaining context information even during complete scene occlusion or when the hand temporarily leaves the field of view.

The resilience of our system to hand-object occlusions was achieved thanks to a custom dataset that we gathered to train the vision module and constitutes one of the key enabling contributions of this study. Looking forward, expanding our dataset to include a broader range of tools and scenarios will enhance the system's generalization, enabling more flexible adaptive assistance. Each company could expand our open-source released dataset with its own tools to tailor the system to its specific needs. Moreover, integrating advanced weight-estimation models or more sophisticated vision algorithms could enhance predictive capabilities and minimize residual errors. For example, our perception pipeline could be extended to estimate object size and assign weights specific to different instances within the same category.

In this study, we evaluated the system using an experimental protocol focused on hand-held industrial tools. The maximum handled weight was approximately 1.7 kg, reflecting the operational range chosen for the experimental

protocol; future work will extend the evaluation to heavier and bulkier objects.

All experiments were conducted in a laboratory environment with lighting conditions configured to approximate those of typical indoor industrial workplaces, ensuring controlled and repeatable testing. Outdoor operation under direct sunlight was beyond the scope of the present study and will be investigated in future studies.

VIII. Conclusion

This work presented the first fully vision-driven control strategy for a bimanual soft upper-limb exoskeleton, enabling real-time, context-aware assistance during industrial tool manipulation. By combining tool recognition, hand tracking with gesture detection, and a fusion layer for reliable context interpretation, the system provides adaptive assistance without requiring obtrusive physiological or hand-mounted sensors.

Experiments with human participants demonstrated significant ergonomic benefits, including reductions of more than 50% in biceps activation compared to the no-support condition.

A key enabling factor was the introduction of an occlusion-robust dataset of annotated industrial tools, which not only improved segmentation accuracy but also could lay the foundation for broader adoption.

Overall, this study establishes vision-based control as a scalable and practical path for advancing exoskeletons from simple supportive devices toward intelligent, adaptive wearable robots. By enhancing ergonomics and mitigating musculoskeletal risks, the proposed approach contributes to safer and more productive industrial workplaces, paving the way for the next generation of context-aware exoskeletons.

References

[1] M. Vukobratovic, "When were active exoskeletons actually born?," *I. J. Humanoid Robotics*, vol. 4, pp. 459–486, 09 2007.

[2] R. S. Mosher, "Handyman to hardiman," *SAE Transactions*, vol. 76, pp. 588–597, 1968.

[3] A. Golabchi, A. Chao, and M. Tavakoli, "A systematic review of industrial exoskeletons for injury prevention: Efficacy evaluation metrics, target tasks, and supported body postures," *Sensors*, vol. 22, no. 7, p. 2714, 2022.

[4] A. Baldassarre, L. G. Lulli, F. Cavallo, L. Fiorini, A. Mariniello, N. Mucci, and G. Arcangeli, "Industrial exoskeletons from bench to field: Human-machine interface and user experience in occupational settings and tasks," *Frontiers in Public Health*, vol. 10, p. 1039680, 2022.

[5] A. Cardoso, A. Ribeiro, P. Carneiro, and A. Colim, "Evaluating exoskeletons for wmsd prevention: A systematic review of applications and ergonomic approach in occupational settings," *International Journal of Environmental Research and Public Health*, vol. 21, no. 12, p. 1695, 2024.

[6] A. Ali, V. Fontanari, W. Schmoelz, and S. K. Agrawal, "Systematic review of back-support exoskeletons and soft robotic suits," *Frontiers in Bioengineering and Biotechnology*, vol. 9, p. 765257, 2021.

[7] M. Xiloyannis, R. Alicea, A. M. Georgarakis, F. L. Haufe, P. Wolf, L. Masia, and R. Riener, "Soft robotic suits: State of the art, core technologies, and open challenges," *IEEE Transactions on Robotics*, vol. 38, no. 3, pp. 1343–1362, 2022.

[8] S. Crea, P. Beckerle, M. De Looze, K. De Pauw, L. Grazi, T. Kermavnar, J. Masood, L. W. O'Sullivan, I. Pacifico, C. Rodriguez-Guerrero, *et al.*, "Occupational exoskeletons: A roadmap toward large-scale adoption. methodology and challenges of bringing exoskeletons to workplaces," *Wearable Technologies*, vol. 2, p. e11, 2021.

[9] J. Chung, D. A. Quirk, M. Applegate, M. Rouleau, N. Degenhardt, I. Galiana, D. Dalton, L. N. Awad, and C. J. Walsh, "Lightweight active back exosuit reduces muscular effort during an hour-long order picking task," *Communications Engineering*, vol. 3, no. 1, p. 35, 2024.

[10] C. Yan, J. J. Banks, B. T. Allaire, D. A. Quirk, J. Chung, C. J. Walsh, and D. E. Anderson, "Musculoskeletal models determine the effect of a soft active exosuit on muscle activations and forces during lifting and lowering tasks," *Journal of Biomechanics*, vol. 176, p. 112322, 2024.

[11] A. Moya-Esteban, M. I. Refai, S. Sridar, H. van der Kooij, and M. Sartori, "Soft back exosuit controlled by neuro-mechanical modeling provides adaptive assistance while lifting unknown loads and reduces lumbosacral compression forces," *Wearable Technologies*, vol. 6, p. e9, 2025.

[12] H. van der Kooij, E. H. van Asseldonk, M. Sartori, C. Basla, A. Esser, and R. Riener, "Ai in therapeutic and assistive exoskeletons and exosuits: Influences on performance and autonomy," *Science Robotics*, vol. 10, no. 104, p. eadt7329, 2025.

[13] D. Hochreiter, K. Schmermbeck, M. Vazquez-Pufleau, and A. Ferscha, "Intention prediction for active upper-limb exoskeletons in industrial applications: A systematic literature review," *Sensors*, vol. 25, no. 17, 2025.

[14] L. Botti and R. Melloni, "Occupational exoskeletons: Understanding the impact on workers and suggesting guidelines for practitioners and future research needs," *Applied Sciences*, vol. 14, no. 1, p. 84, 2024.

[15] F. Missiroli, N. Lotti, M. Xiloyannis, L. H. Sloot, R. Riener, and L. Masia, "Relationship between muscular activity and assistance magnitude for a myoelectric model based controlled exosuit," *Frontiers in Robotics and AI*, vol. 7, December 2020.

[16] M. Sierotowicz, D. Brusamento, B. Schirrmeister, M. Connan, J. Bornmann, J. Gonzalez-Vargas, and C. Castellini, "Unobtrusive, natural support control of an adaptive industrial exoskeleton using force myography," *Frontiers in Robotics and AI*, vol. 9, p. 919370, 2022.

[17] O. Sherif, M. M. Bassuoni, and O. Mehrez, "A survey on the state of the art of force myography technique (fmg): Analysis and assessment," *Medical & Biological Engineering & Computing*, vol. 62, no. 5, pp. 1313–1332, 2024.

[18] N. J. Jarque-Bou, J. L. Sancho-Bru, and M. Vergara, "A systematic review of emg applications for the characterization of forearm and hand muscle activity during activities of daily living: Results, challenges, and open issues," *Sensors*, vol. 21, no. 9, p. 3035, 2021.

[19] J.-H. Sul, L. Piyathilaka, D. Moratuwage, S. Dunu Arachchige, A. Jayawardena, G. Kahandawa, and D. M. G. Preethichandra, "Electromyography signal acquisition, filtering, and data analysis for exoskeleton development," *Sensors*, vol. 25, no. 13, p. 4004, 2025.

[20] M. Abdoli-Eramaki, C. Damecour, J. Christenson, and J. Stevenson, "The effect of perspiration on the semg amplitude and power spectrum," *Journal of Electromyography and Kinesiology*, vol. 22, pp. 908–913, Dec. 2012. Epub 2012 May 19.

[21] E. Mobedi, G. Solak, and A. Ajoudani, "A framework for adaptive load redistribution in human-exoskeleton-cobot systems," *IEEE Robotics and Automation Letters*, vol. 10, no. 6, pp. 5927–5934, 2025.

[22] E. Tricomi, G. Piccolo, F. Russo, X. Zhang, F. Missiroli, S. Ferrari, L. Gionfrida, F. Ficuciello, M. Xiloyannis, and L. Masia, "Leveraging geometric modeling-based computer vision for context aware control in a hip exosuit," *IEEE Transactions on Robotics*, 2025.

[23] C. Wang, Z. Pei, Y. Fan, S. Qiu, and Z. Tang, "Review of vision-based environmental perception for lower-limb exoskeleton robots," *Biomimetics*, vol. 9, no. 4, p. 254, 2024.

[24] F. Missiroli, P. Mazzoni, N. Lotti, E. Tricomi, F. Braghin, L. Roveda, and L. Masia, "Integrating computer vision in exosuits for adaptive support and reduced muscle strain in industrial environments," *IEEE Robotics and Automation Letters*, vol. 9, no. 1, pp. 859–866, 2024.

[25] N. Bernstein, "The coordination and regulation of movements," *(No Title)*, 1967.

[26] J. Kuhn, J. Ringwald, M. Schappler, L. Johannsmeier, and S. Haddadin, "Towards semi-autonomous and soft-robotics enabled upper-limb exoprosthetics: First concepts and robot-based emulation prototype," in *2019 International Conference on Robotics and Automation (ICRA)*, pp. 9180–9186, 2019.

[27] F. Hundhausen, S. Hubschneider, and T. Asfour, "Grasping with humanoid hands based on in-hand vision and hardware-accelerated cnns," in *2023 IEEE-RAS 22nd International Conference on Humanoid Robots (Humanoids)*, pp. 1–7, 2023.