Journal of Hunan University

(Natural Sciences)

Vol. 52 No. 7

July 2025

Available online at

https://jonuns.com

Open Access Article

https://doi.org/10.55463/issn.1674-2974.52.7.2

Assistive Robot for Automatic Sorting Using Voice-Guided Human-Machine Interaction

Robinson Jiménez-Moreno[1*], Ricardo A. Castillo[1], Alexander Aponte-Moreno[1]

[1]Professors of Department of Mechatronics, Engineering Faculty, Universidad Militar Nueva Granada, Bogotá, Colombia

* Corresponding author: robinson.jimenez@unimilitar.edu.co

Article History:
Received: June 11, 2025
Revised: July 18, 2025
Accepted: July 28, 2025
Published: August 30, 2025

Abstract: This paper presents the development of a virtual environment for object classification and location by an assistive robot using voice-guided human-machine interaction to facilitate the storage or ordering of products by means of robotic systems, reducing potential risks due to excess load or repeatability to human operators. The proposed approach integrates computer vision techniques and transformer-based speech recognition models within a virtual simulation environment. Specifically, a ResNet18 neural network, selected for its low computational demand and high efficiency in classification and localization tasks, is used to accurately identify objects. As a contribution to the state of the art, a human-machine interaction environment is developed with natural language processing algorithms oriented toward industrial applications, where the sorting order is specified by voice commands captured and transcribed by a wav2vec-based speech-to-text algorithm, allowing users to interact naturally and efficiently with the robotic system. Experimental validation demonstrates the robustness of object detection and the reliability of speech recognition, highlighting the system's effectiveness and potential applications in automated industrial scenarios.

Keywords: ResNet, Speech to text (STT), Human Robot collaboration, Automation.

利用语音引导的人机交互技术实现自动分类的辅助机器人

摘要： 本文介绍了一个虚拟环境的开发情况，该环境用于辅助机器人利用语音引导的人机交互进行物体分类和定位。该方法在虚拟仿真环境中集成了计算机视觉技术和基于变压器的语音识别模型。具体来说，ResNet18 神经网络因其在分类和定位任务中的低计算需求和高效率而被选中，用于准确识别物体。排序顺序由基于 wav2vec 的语音到文本（STT）算法捕获和转录的语音命令指定，从而使用户能够与机器人系统进行自然、高效的交互。实验验证证明了物体检测的鲁棒性和语音识别的可靠性，突出了该系统的有效性及其在自动化工业场景中的潜在应用。

关键词：ResNet、语音转文本 STT、人机协作、自动化。

1. Introduction

The transition from Industry 4.0 to 5.0 [1] has significantly impacted manufacturing processes through artificial intelligence (AI) integration [2], showcasing numerous advancements, such as digital twins [3] and their combination with large language models [4], to validate processes in simulation environments.

Within AI-driven manufacturing processes, those supported by computer vision hold great relevance, emphasizing the importance of dataset quality, various existing models, and their comparative analysis [5]. Prominent AI algorithms designed for image feature extraction include Faster RCNN [6] and ResNet [7], which have industrial applications in aquaculture [8] and agriculture [9]. These techniques predominantly focus on image analysis [10], which facilitates both classification and localization tasks [11].

The exponential evolution of AI has further led to the integration of these models with advanced network architectures, such as transformer networks. These integration have demonstrated notable applications in railway failure prediction [12], electromagnetic tomography image reconstruction [13], and vehicle intrusion detection [14]. Transformer architectures particularly excel in speech-based language models [15], encompassing audiovisual representation [16], speech-based detection of health conditions [17], dangers like intoxication [18], verbal threats [19], and general emotion recognition [20].

Robot control based on voice commands has also been effectively implemented, including applications in robotic soccer [21], emotion recognition [22], educational [23-24], music-oriented robots for classrooms, and broader human-robot interaction scenarios [25]. In the Industry 4.0 and 5.0 context, voice-controlled robotics have significantly advanced the automation of human-machine interaction, leveraging transformer-based models, such as wav2vec [26], which are recognized for their efficiency.

Therefore, combining image classification and localization models with voice recognition technologies, this paper presents the design and implementation of a virtual environment for product sorting based on voice-driven human-robot interaction. Objects for sorting are identified using a ResNet18 network due to its predefined architecture, which is computationally efficient for object classification and localization tasks. Order recognition and validation are performed using a speech-to-text (STT) algorithm based on wav2vec [27], enabling users to vocally specify objects assigned to each row of a 6x6 array. Using the aforementioned artificial intelligence algorithms, we contribute to the state of the art with a human-machine interaction environment based on natural language processing communication and oriented to industrial applications that facilitate the performance of activities by the operator.

The rest of the article is structured as follows: the next section describes the methodology employed, detailing the integration process using MATLAB and Coppelia simulation environments. This is followed by a discussion of the experimental results obtained. Finally, conclusions are drawn highlighting the effectiveness and challenges of the proposed approach.

2. Methodology

The proposed development employs an applied research methodology derived from the approach to solving a specific problem in an industrial environment, which facilitates the storage or ordering of products by means of robotic systems, reducing potential risks due to excessive load or repeatability to human operators. The automatic sorting process begins with the recognition of six predefined objects: spheres, cubes, cylinders, stars, a pink box, and a black box. To minimize handling time, a transfer learning approach with a limited class count is employed, using a ResNet18 model [28].

Figure 1 demonstrates samples from the training database, comprising images of objects arranged

differently within a 6x6 array, facilitating object type localization.

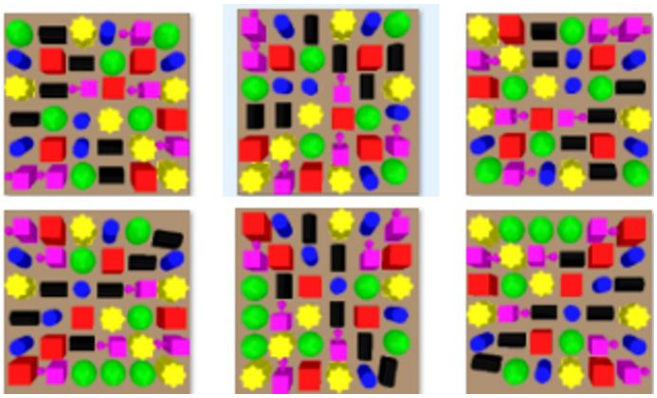

Figure 1. Sample training images (source: authors)

The training parameters detailed in Table 1 guide the model development.

Table 1. Training parameters (source: developed by the authors)

Parameter	Value
Input volume	600X600X3
Optimizer	ADAM
Number of Epochs	50
Minibathc size	2
Learning rate	1*10-6

The learning curve, depicted in Figure 2, reveals an accuracy peak of 92.7%, stabilizing around the 20th epoch.

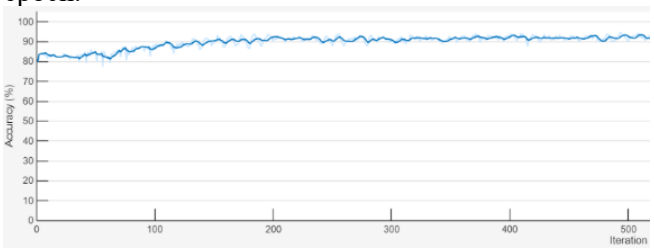

Figure 2. Sample training images (source: authors)

Figure 3 validates the trained network's ability to accurately classify and localize various objects, displaying precise bounding boxes crucial for robotic gripping.

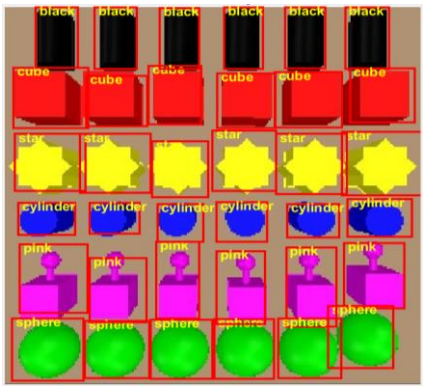

Figure 3. Classification and validation of localization (Source: developed by the authors)

The wav2vec algorithm [27] conducts audio recognition and converts it to text. This self-supervised approach creates speech representations from raw audio inputs without the need for traditional preprocessing steps such as cepstral coefficients or spectrogram generation [29].

Wav2Vec employs a multilayer CNN [30] to transform input waveforms into low-dimensional latent representations. Subsequently, a transformer network [31] captures contextual dependencies between audio segments.

Table 2 lists the audio capture parameters used to recognize object labels, enabling voice-driven interaction within the Coppelia simulation environment.

Table 2. Voice recording (source: authors)

Parameter	Setting Value
Recording time	3 seg
Sample frequency	22050
Number of bits	16
Number of channels used	1 (monophonic)

Integration is achieved through a MATLAB-based environment (Figure 4), connecting users to the Coppelia simulator via the STAR CONNECTION button (green button). Automatic sorting starts with the START SORTING button (aquamarine button). The process can be stopped using the STOP button (red button) from the interface. Additionally, a user help button is included to visualize the action to be taken by each button by means of a pop-up text. Before starting the sorting, users must indicate the location of the objects in the columns using voice commands; for this purpose, an intuitive button interface is available.

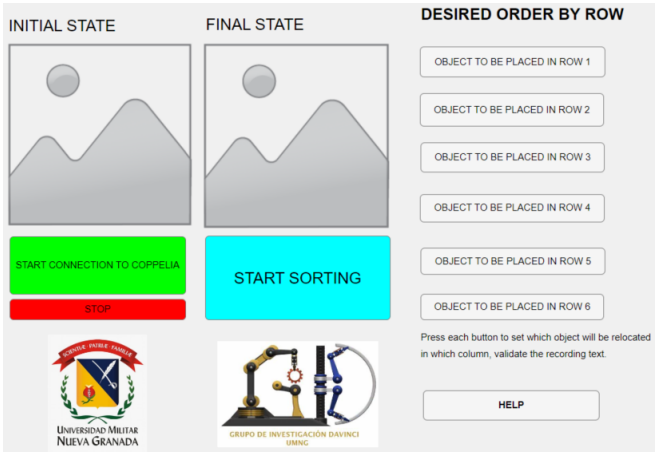

Figure 4. Operating environment from MATLAB (source: authors)

Figure 5 illustrates the Coppelia environment where the user and the robot with the objects to be organized are simulated. There are two vision sensors: one for object identification and the other for the arrangement structure that corresponds to the 6x6 boxes that will contain each object.

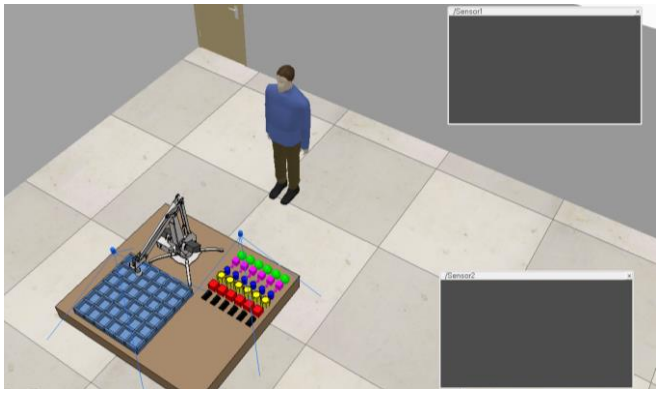

Figure 5. Coppelia work environment (Source: authors)

Figure 6 illustrates the robotic arm used in the virtual environment and the associated degrees of freedom that allow the inference of the kinematic equations of motion for robot control.

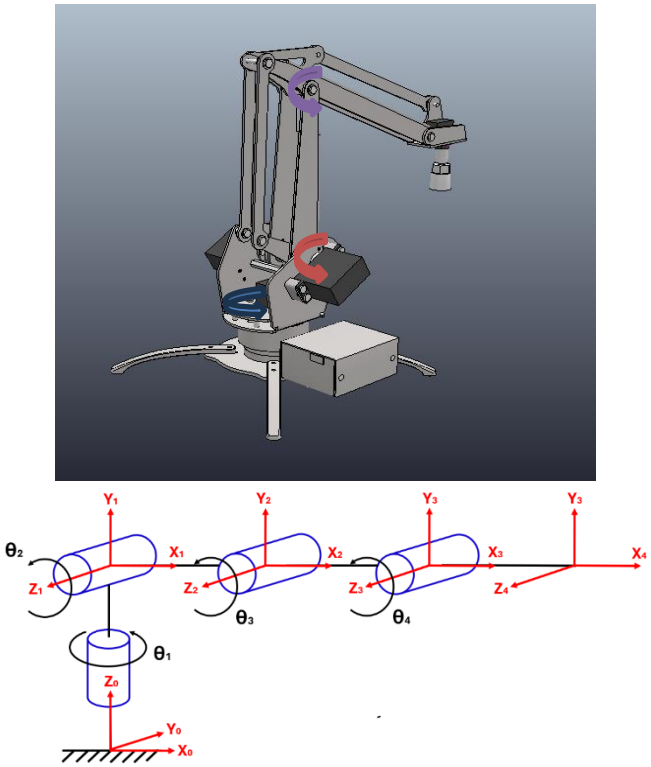

Figure 6. Serial robot manipulator schematic and DH coordinate systems (source: authors)

3. Results

To perform direct kinematic modeling, the Denavit-Hartenberg (DH) convention is adopted, the serial robot parameters are listed in Table 3. Homogeneous transformation matrices for each robotic arm link were calculated using trigonometric identities to obtain complete direct kinematics calculation.

Table 2. Voice recording (Source: developed by the authors)

i	θ_i	d_i	a_i	α_i	Description
1	θ_1	d_1	0	$\pi/2$	Transformation from S0 to S1
2	θ_2	0	a_2	0	Transformation from S1 to S2
3	θ_3	0	a_3	0	Transformation from S2 to S3 is shown.
4	$-(\theta_2+\theta_3+\pi/2)$	0	a_4	0	Orientation of the final effector to vertical orientation

The link-specific homogeneous transformation matrices are calculated using the trigonometric identity for negative rotation.

$$A^{01} = \begin{bmatrix} cos(\theta_1) & 0 & sin(\theta_1) & 0 \\ sin(\theta_1) & 0 & -cos(\theta_1) & 0 \\ 0 & 1 & 0 & d_1 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad (1)$$

$$A_1{}^2 = \begin{bmatrix} cos(\theta_2) & -sin(\theta_2) & 0 & a_2 * cos(\theta_2) \\ sin(\theta_2) & cos(\theta_2) & 0 & a_2 * sin(\theta_2) \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad (2)$$

$$A_2{}^3 = \begin{bmatrix} cos(\theta_3) & -sin(\theta_3) & 0 & a_3 * cos(\theta_3) \\ sin(\theta_3) & cos(\theta_3) & 0 & a_3 * sin(\theta_3) \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad (3)$$

$$A_3{}^4 = \begin{bmatrix} -sin(\theta_2+\theta_3) & cos(\theta_2+\theta_3) & 0 & -a_4 * sin(\theta_2+\theta_3) \\ -cos(\theta_2+\theta_3) & -sin(\theta_2+\theta_3) & 0 & -a_4 * cos(\theta_2+\theta_3) \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad (4)$$

The complete transformation from the base system to the final effector (frame 4) is obtained by multiplying the following matrices:

$$A_0{}^4 = A_0{}^1 \cdot A_1{}^2 \cdot A_2{}^3 \cdot A_3{}^4 \quad (5)$$

This matrix gives the end effector's position and orientation with respect to the base system. It can be expressed as follows:

$$A^{04} = \begin{bmatrix} 0 & cos(\theta_1) & sin(\theta_1) & \cos(\theta_1) * [a_3 * cos(\theta_2+\theta_3) + a_2 * \cos(\theta_2)] \\ 0 & sin(\theta_1) & -cos(\theta_1) & \sin(\theta_1) * [a_3 * cos(\theta_2+\theta_3) + a_2 * \cos(\theta_2)] \\ -1 & 0 & 1 & d_1 - a_4 + a_3 * sin(\theta_2+\theta_3) + a_2 * \sin(\theta_2) \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad (6)$$

where

$$R^{04} = \begin{bmatrix} 0 & cos(\theta_1) & sin(\theta_1) \\ 0 & sin(\theta_1) & -cos(\theta_1) \\ -1 & 0 & 1 \end{bmatrix} \quad (7)$$

Is the rotation matrix from S0 to S4, and the fourth column of the total transformation matrix is the EEF position (P04).

$$x = \cos(\theta_1) * [a_3 * cos(\theta_2+\theta_3) + a_2 * \cos(\theta_2)] \quad (8)$$

$$y = \sin(\theta_1) * [a_3 * cos(\theta_2+\theta_3) + a_2 * \cos(\theta_2)] \quad (9)$$

$$z = d_1 - a_4 + a_3 * sin(\theta_2+\theta_3) + a_2 * \sin(\theta_2) \quad (10)$$

where d_1 = 8.951, a_2 = 14.8, a_3 = 16 cm, and a_4 = 1.776 cm.

Algorithm integration validation occurs within the MATLAB application, initiating connections to Coppelia. Users can visualize initial object placements

and then sequentially press interface buttons to vocally designate objects per row. Upon pressing "start sorting," robotic gripping and object rearrangement begin (Figure 7), confirming accurate classification and localization.

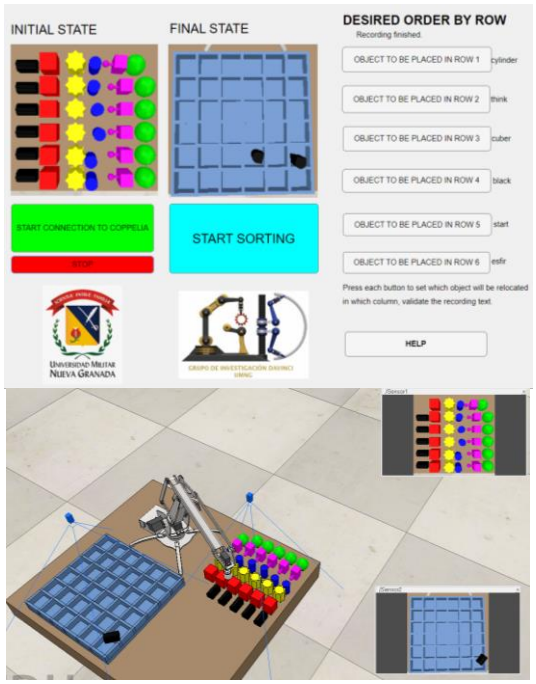

Figure 7. Matlab application commanding robotic manipulator in Coppelia simulator (source: authors)

Figure 8 illustrates the robotic sorting progress, showing the correct classification and localization performed by ResNet. A few gripping issues with the suction-based final effector are evidenced in some tests. Objects sometimes fall or rotate incorrectly due to simulated gravitational effects, occasionally disrupting precise alignment.

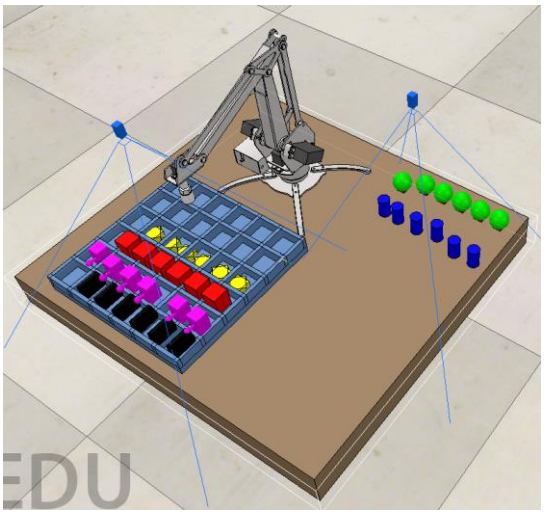

Figure 8. Sorting progress in the Coppelia virtual environment (Source: authors)

Clear vocal pronunciation is crucial for accurate recording. Misrecognitions can be corrected by repeating vocal commands through the interface. Figure 9 shows the user assistance messages accessible via the help button.

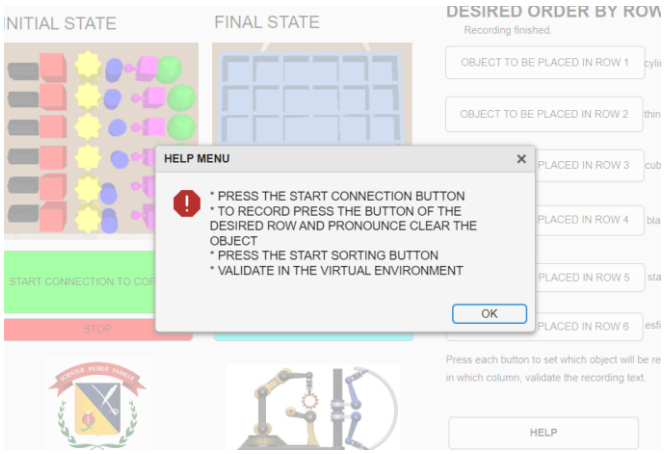

Figure 9. Help button (Source: developed by the authors)

4. Discussion

Using the Wav2Vec algorithm, the results obtained for human-robot interaction by voice are optimized for the results obtained in previous works using audio preprocessing and convolutional network classification, such as the one presented in [32]. Although voice detection is more fluent, the need for clear pronunciation was evident given that the researchers' mother tongue is Spanish. It is by extension an improvement to the production line processes the use of robots in the process, but their collaborative work is facilitated and can be improved with a communicative interaction as natural as possible.

As recommendations and future work, a bidirectional communication based on speech-to-text transcription (STT) algorithms can be handled, for example, using the Whisper tool developed by OpenAI, so that the robotic arm confirms the desired order or reports the task completion.

5. Conclusion

The wav2vec-based STT model significantly improves the efficiency of voice interaction, optimizing robotic control without implementation complications such as those based on audio preprocessing.

The integration of ResNet's robotic kinematics and precise object localization contributes to the success of object repositioning. Adjusting detection thresholds for confidence levels and the dimensions of the bounding box where the object is located is essential for effective object grasping and repositioning.

The integration of algorithms allowed the validation of a friendly and easy to operate human-robot interaction interface. Future work includes the start of the process by voice and not by help buttons, adjusting the code to noise thresholds that activate the voice capture and subsequent associated execution

Declarations

Author Contributions

Conceptualization, formal analysis, and writing—review and editing R. Jiménez-Moreno Castillo Ricardo and Aponte Alexander.; methodology, validation, investigation supervision, project administration, funding acquisition, and data curation Jiménez-Moreno R.; Software, writing—original draft preparation and visualization Castillo Ricardo and Aponte Alexander. All authors have read and approved the published version of the manuscript.

Funding

Product derived from the research project titled " Design of a human-robot interaction model using deep learning algorithms" INV-ING-3971, which was financed by the Vice-Rector for Research of the Universidad Militar Nueva Granada in 2024.

Acknowledgements

The authors would like to thank the Universidad Militar Nueva Granada for the time and resources available for the development of this article.

Statement of the Institutional Review Board

The study was conducted in accordance with the Declaration of Helsinki and approved by the Ethics Committee of the Universidad Militar Nueva Granada (project INV-ING-3971, date of approval: January 18, 2024).

Conflicts of Interest

The author declares that there are no conflicts of interest regarding the publication of this manuscript. In addition, the authors have completely observed the ethical issues, including plagiarism, informed consent, misconduct, data fabrication and/or falsification, double publication and/or submission, and redundancies.

References

[1] L. Monferdini, L. Tebaldi, and E. BOTTANI. Industry 4.0 to Industry 5.0: Opportunities, Challenges, and Future Logistics Perspectives *Procedia Comput Sci*, 2025, 253: 2941-2950. doi: 10.1016/j.procs.2025.02.018

[2] M. E. LATINO A maturity model for assessing the implementation of Industry 5.0 in manufacturing SMEs: learning from theory and practice *Technological Forecasting and Social Change*, 2025, 214: 124045. https://doi.org/10.1016/j.techfore.2025.124045

[3] A. Massaro, F. Santarsiro, G. Schiuma, Advanced electronic controller circuits enabling production processes and AI-driven KM in Industry 5.0. *Journal of Industrial Information Integration*, 2025, 45: 100841, https://doi.org/10.1016/j.jii.2025.100841https://doi.org/10.1016/j.jii.2025.100841

[4] C. CHEN, K. ZHAO, J. LENG, C. LIU, J. FAN, P. ZHENG. Integrating the large language model and digital twins in the context of Industry 5.0: Framework, challenges, and opportunities *Robotics and Computer-Integrated Manufacturing*, 2025, 94: 102982. https://doi.org/10.1016/j.rcim.2025.102982

[5] J. LI, X. HU, A. LUCIC, Y. WU, I.C.F.S. CONDOTTA, R. N. DILGER, N. AHUJA, and A. R. GREEN-MILLER, "Promote computer vision applications in pig farming scenarios: high-quality dataset, fundamental models, and comparable performance," Journal of Computer Vision, vol. *Journal of Integrative Agriculture*, 2024, https://doi.org/10.1016/j.jia.2024.08.014

[6] B. HELIAN, X. HUANG, M. YANG, Y. BIAN, and M. GEIMER Estimation of excavator bucket fill using a computer vision-based depth map and faster R-CNN *Automation in Construction*, 2024, 166: 105592, https://doi.org/10.1016/j.autcon.2024.105592

[7] J. Liao, L. Guo, L. Jiang, C. Yu, W. Liang, K. Li, and F. Po A machine learning-based feature extraction method for image classification using ResNet architecture is proposed. *Digital Signal Processing*, 2025, 160: 105036. doi: 10.1016/j.dsp.2025.105036

[8] H. Yu, H. Song, L. Xu, D. Li, Y. Chen, SED-RCNN-BE: A SE-Dual channel RCNN network optimized binocular estimation model for automatic size estimation of free swimming fish in aquaculture, J. Phys. Res. Commun. *Expert Systems with Applications*, 2024, 255(Part A): 124519, https://doi.org/10.1016/j.eswa.2024.124519

[9] Z. Ren, F. Tian, S. Wang, S. Chen, Research on maize leaves surface action potential recognition method based on ResNet-18SE. *Smart Agricultural Technology*, 2025, 10: 100819. https://doi.org/10.1016/j.atech.2025.100819.

[10] W. Du, M. Qian, S. He, L. Xu, X. Zhang, M. Huang, N. Chen. Improved ResNet method for urban flooding water depth estimation from social media images *Measurement*, 2025, 242(Part D): 116114. doi: 10.1016/j.measurement.2024.116114

[11] A. KHATTAK, P. W. CHAN, F. CHEN, A. H. ALMALIKI. Deep ResNet Strategy for Classifying Wind Shear Intensity Near Airport Runway *Computer Modeling in Engineering and Sciences*, 2025, 142(2): 1565-1584. https://doi.org/10.32604/cmes.2025.059914https://doi.org/10.32604/cmes.2025.059914

[12] X. Wang, J. Dai, X. Liu. A spatial-temporal neural network based on ResNet-Transformer for predicting railroad broken rails. *Advanced Engineering Informatics*, 2025, 65(Part A): 103126. https://doi.org/10.1016/j.aei.2025.103126https://doi.org/10.1016/j.aei.2025.103126

[13] XI. Liu, H. Feng, Y. Wang, D. Li, K. Zhang. Hybrid ResNet and transformer model for efficient image reconstruction of electromagnetic tomography *Flow Measurement and Instrumentation*, 2025, 102: 102843. doi: 10.1016/j.flowmeasinst.2025.102843

[14] Z. Wu, M. Li. ResNet-Swin Transformer based intrusion detection system for in-vehicle network. *Expert Systems with Applications*, 2025, 127547, https://doi.org/10.1016/j.eswa.2025.127547

[15] L. F. Parra-Gallego, T. Arias-Vergara, J. R. Orozco-Arroyave. Multimodal evaluation of voicemail customer satisfaction using speech and language representations *Digital Signal Processing*, 2025, 156(Part B): 104820. DOI: 10.1016/j.dsp.2024.104820

[16] J. X. Zhang, G. WAN, J. GAO, Z. H. Ling, Audio-visual representation learning via knowledge distillation from speech foundation models, Applied Speech Science, 89, e013–e018. *Pattern Recognition*, 2025, 162:111432. doi: 10.1016/j.patcog.2025.111432
[17] S. AUROBINDO, R. PRAKASH, M. RAJESHKUMAR. Comparative analysis of different time-frequency image representations for the detection and severity classification of dysarthric speech using deep learning (DL) *Results in Engineering,* 2025, 25: 104561. https://doi.org/10.1016/j.rineng.2025.104561
[18] A. Albuquerque, S. Chibuoyim Uche, E. Agu, Intoxication detection from speech using representations learned from self-supervised pre-training. *Smart Health*, 2025, 100562. https://doi.org/10.1016/j.smhl.2025.100562
[19] T. NEUMAIER. The representation of threatening speech in Late Modern English trials *Journal of Pragmatics*, 2025, 237: 55-67. https://doi.org/10.1016/j.pragma.2025.01.004https://doi.org/10.1016/j.pragma.2025.01.004
[20] A. Chakhtouna, S. Sekkate, A. Abib. Modeling Speech Emotion Recognition using ImageBind representations *Procedia Comput Sci*, 2024, 236: 428-435. https://doi.org/10.1016/j.procs.2024.05.050
[21] P. FIATI. SMILE: A verbal and graphical user interface tool for speech-control of soccer robots in Ghana. *Cognitive Robotics*, 2021, 1: 25-28. https://doi.org/10.1016/j.cogr.2021.03.001
[22] X. Kang. Speech emotion recognition algorithm of intelligent robot based on ACO-SVM. *International Journal of Cognitive Computing in Engineering*, 2025, 6: 131-142. https://doi.org/10.1016/j.ijcce.2024.11.008https://doi.org/10.1016/j.ijcce.2024.11.008
[23] X. Zhou. Application of entertainment performance robots in a music network classroom based on speech sensor recognition and artificial intelligence *Entertainment Computing*, 2025, 52:100782. https://doi.org/10.1016/j.entcom.2024.100782
[24] Z. YING Experience of an intelligent speech robot in an online music classroom based on deep learning and virtual reality *Entertainment Computing,* 2025, 52: 100795. doi: 10.1016/j.entcom.2024.100795
[25] N. GRÁGEDA, C. Busso, E. Avarrado, R. GARCÍA, R. Muru, F. Huenupan, N. Benecerra Yoma. Speech emotion recognition in real static and dynamic human-robot interaction scenarios *Computer Speech & Language*, 2025, 89: 101666. doi: 10.1016/j.csl.2024.101666
[26] S. Park, the M. Mark, B. Park, H. Hong. Speaker-Specific Emotion Representations in Wav2vec 2.0-Based Modules for Speech Emotion Recognition *Computers, Materials and Continua*, 2023, 77(1): 1009-1030. https://doi.org/10.32604/cmc.2023.041332https://doi.org/10.32604/cmc.2023.041332
[27] A. Baevski, H. Zhou, A. Mouhamed, M. Aulique. Wav2vec 2.0: A Framework for the Self-Supervised Learning of Speech Representations *Computer Science, Computation and Language,* 2020. https://doi.org/10.48550/arXiv.2006.11477.
[28] F. NAVEED, A. MASIH, J. MAHMOOD, M. AHMED, A. ALI, A. SADDIQA, M. S. HAMZA ABDULNABI, and E. AGBOZO. Sustainable AI for plant disease classification using ResNet18 in few-shot learning. *Array*, 2025, 26: 100395. https://doi.org/10.1016/j.array.2025.100395
[29] R. JIMÉNEZ MERNONO, A. A. ESPITIA CUBILLOS, E. RODRÍGUEZ CARMONA. Interactive communication human-robot interface for reduced mobility people assistance. *IAES International Journal of Artificial Intelligence*, 2025, 14(2). http://doi.org/10.11591/ijai.v14.i2.pp917-924
[30] Y. LECUN, L. BOTTOU, Y. BENGIO, P. HAFFNER. Gradient-Based Learning for Document Recognition *Proceedings of the IEEE*, 1998, 86(11): 2278-2324. https://doi.org/10.1109/5.726791
[31] A. VASWANI, N. SHAZEER, N. PARMAR, J. USZKOREIT, L. JONES, A. N. GOMEZ, Ł. KAISER and POLOSUKHIN. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems.* 6000–6010, 2017. Curran Associates Inc., Red Hook, NY, USA,
[32] R. JIMÉNEZ-MORENO and R. A. CASTILLO Deep learning speech recognition for a residential assistant robot *International Journal of Artificial Intelligence*, 2022, 12(2): 585-592. http://doi.org/10.11591/ijai.v12.i2.pp585-592

参考文:

[1] L. Monferdini, L. Tebaldi, and E. BOTTANI. 从工业 4.0 到工业 5.0：物流领域的机遇、挑战和未来展望。*Procedia 计算机科学*, 2025, 253: 2941-2950. https://doi.org/10.1016/j.procs.2025.02.018
[2] M. E. LATINO. 评估制造业中小企业实施工业 5.0 的成熟度模型：从理论和实践中学习。*技术预测与社会变革*, 2025, 214: 124045. doi: 10.1016/j.techfore.2025.124045
[3] A. Masaso, F. Santarsiro, G. SCHIUMA, 先进的电子控制电路助力工业 5.0 中的生产流程和人工智能驱动的知识管理。*工业信息集成杂志*, 2025, 45: 100841, https://doi.org/10.1016/j.jii.2025.100841
[4] C. Chen, K. Zhao, J. Leng, C. Liu, J. Fan, P. ZHENG. 工业 5.0 背景下大型语言模型与数字孪生的集成：框架、挑战与机遇。*机器人与计算机集成制造,* 2025, 94: 102982. https://doi.org/10.1016/j.rcim.2025.102982
[5] J. LI, X. HU, A. LUCIC, Y. WU, I.C.F.S. CONDOTTA, R. N. DILGER, N. AHUJA, A. R. GREEN-MILLER, 推动计算机视觉在养猪场景中的应用：高质量数据集、基础模型和可比性能。*综合农业杂志*, 2024, https://doi.org/10.1016/j.jia.2024.08.014
[6] B. HELIAN, X. HUANG, M. YANG, Y. BIAN, M. GEIMER. 基于计算机视觉的挖掘机铲斗填充量估算，使用深度图和 Faster R-CNN。 *建筑自动化*, 2024, 166: 105592, https://doi.org/10.1016/j.autcon.2024.105592
[7] J. Liao, L. Guo, L. Jiang, C. Yu, W. Liang, K. Li, F. POP. 一种基于机器学习、使用 ResNet 架构进行图像分类的特征提取方法。*数字信号处理*, 2025, 160: 105036. https://doi.org/10.1016/j.dsp.2025.105036
[8] H. Yu, H. Song, L. Xu, D. Li, Y. Chen, SED-RCNN-BE：基于 SE-双通道 RCNN 网络优化的双目估计模型，用于水产养殖中自由游动鱼类的自动

尺寸估计。*专家系统及其应用*, 2024, 255(Part A): 124519, https://doi.org/10.1016/j.eswa.2024.124519
[9] Z. Ren, F. Tian, S. Wang, S. CHEN. 基于 ResNet-18SE 的玉米叶片表面动作电位识别方法研究。*智慧农业技术*, 2025, 10: 100819. https://doi.org/10.1016/j.atech.2025.100819.
[10] W. Du, M. Qian, S. He, L. Xu, X. Zhang, M. Huang, N. CHEN. 一种基于社交媒体图像的城市洪水深度估算的改进型 ResNet 方法。*测量*, 2025, 242(Part D): 116114. https://doi.org/10.1016/j.measurement.2024.116114
[11] A. KHATTAK, P. W. CHAN, F. CHEN, A. H. ALMALIKI. 深度 ResNet 策略用于机场跑道附近风切变强度的分类。*工程与科学中的计算机建模*, 2025, 142(2): 1565-1584. https://doi.org/10.32604/cmes.2025.059914
[12] X. Wang, J. DAI, X. LIU. 基于 ResNet-Transformer 的时空神经网络，用于预测铁路断轨。*《高级工程信息学》*, 2025, 65(Part A): 103126. https://doi.org/10.1016/j.aei.2025.103126
[13] X. Liu, H. Feng, Y. Wang, D. Li, K. Zhang. ResNet 与 Transformer 混合模型，用于高效重建电磁层析成像。*流量测量与仪器*, 2025, 102: 102843. https://doi.org/10.1016/j.flowmeasinst.2025.102843
[14] Z. Wu, M. LI. 基于 ResNet-Swin Transformer 的车载网络入侵检测系统。*专家系统及其应用*, 2025, 127547, https://doi.org/10.1016/j.eswa.2025.127547
[15] L. F. Parra-Gallego, T. Arias-Vergara, J. R. OROZCO-ARROYAVE. 使用语音和语言表征对语音邮件中的客户满意度进行多模式评估。*数字信号处理*, 2025, 156(Part B): 104820. https://doi.org/10.1016/j.dsp.2024.104820
[16] J. X. Zhang, G. Wang, J. GAO, Z.H. LING, 通过语音基础模型中知识提炼进行视听表征学习。*模式识别*, 2025, 162:111432. https://doi.org/10.1016/j.patcog.2025.111432https://doi.org/10.1016/j.patcog.2025.111432
[17] S. AUROBINDO, R. PRAKASH, M. RAJESHKUMAR. 运用深度学习对构音障碍语音进行检测和严重程度分类，对不同的时频图像表征进行比较分析。*工程研究成果*, 2025, 25: 104561. https://doi.org/10.1016/j.rineng.2025.104561
[18] A. Albuquerque, S. Chibuoyim Uche, E. AGU, 使用从自监督预训练中学习到的表征，从语音中检测醉酒状态。*智能健康*, 2025, 100562. https://doi.org/10.1016/j.smhl.2025.100562
[19] T. NEUMAIER. 晚期现代英国审判中威胁性言语的表现。*语用学杂志*, 2025, 237: 55-67. https://doi.org/10.1016/j.pragma.2025.01.004https://doi.org/10.1016/j.pragma.2025.01.004
[20] A. Chakhtouna, S. Sekkate, A. ADIB. 通过 ImageBind 表征建模语音情感识别。*Procedia 计算机科学* 2024, 236: 428-435. https://doi.org/10.1016/j.procs.2024.05.050
[21] P. FIATI. SMILE：加纳足球机器人语音控制的语音和图形用户界面工具。*认知机器人*, 2021, 1: 25-28. https://doi.org/10.1016/j.cogr.2021.03.001
[22] X. KANG. 基于 ACO-SVM 的智能机器人语音情感识别算法。*国际工程认知计算杂志*, 2025, 6: 131-142. https://doi.org/10.1016/j.ijcce.2024.11.008
[23] X. ZHOU. 基于语音传感器识别和人工智能的娱乐表演机器人在音乐网络教室的应用。*娱乐计算*, 2025, 52:100782. https://doi.org/10.1016/j.entcom.2024.100782
[24] Z. YING. 基于深度学习和虚拟现实的智能语音机器人在音乐在线课堂中的应用体验。*娱乐计算,* 2025, 52: 100795. https://doi.org/10.1016/j.entcom.2024.100795https://doi.org/10.1016/j.entcom.2024.100795
[25] N. GRÁGEDA, C. Busso, E. Avarrado, R. GARCÍA, R. Muru, F. Huenupan, N. BECERRA YOMA. 真实静态和动态人机交互场景中的语音情感识别。*计算机语音与语言*, 2025, 89: 101666. https://doi.org/10.1016/j.csl.2024.101666
[26] S. Parker, M. Mark, B. Parker, H. HONG. 在基于 Wav2vec 2.0 的模块中使用说话人特定情感表征进行语音情感识别。*计算机、材料与连续体*, 2023, 77(1): 1009-1030. https://doi.org/10.32604/cmc.2023.041332
[27] A. Baevski, H. Zhou, A. Mouhamed, M. Aulique. Wav2vec 2.0：语音表征的自监督学习框架。*计算机科学、计算与语言*, 2020. https://doi.org/10.48550/arXiv.2006.11477.
[28] F. NAVEED, A. MASIH, J. MAHMOOD, M. AHMED, A. ALI, A. SADDIQA, M. S. HAMZA ABDULNABI, E. AGBOZO. 使用 ResNet18 进行小样本学习，实现可持续的植物病害分类 AI。*阵列*, 2025, 26: 100395. https://doi.org/10.1016/j.array.2025.100395
[29] R. JIMÉNEZ MORENO, A. A. ESPITIA CUBILLOS, E. RODRÍGUEZ CARMONA. 用于行动不便人士协助的交互式人机交互界面。*IAES 国际人工智能期刊*, 2025, 14(2). http://doi.org/10.11591/ijai.v14.i2.pp917-924
[30] Y. LECUN, L. BOTTOU, Y. BENGIO, P. HAFFNER. 基于梯度的学习应用于文档识别。*IEEE 论文集*, 1998, 86(11): 2278-2324. https://doi.org/10.1109/5.726791
[31] A. VASWANI, N. SHAZEER, N. PARMAR, J. USZKOREIT, L. JONES, A. N. GOMEZ, Ł. KAISER, and I. POLOSUKHIN. 你只需要专注。第 *31 届国际神经信息处理系统会议论文集,* 2017, 6000-6010. Curran Associates Inc., Red Hook, NY, USA
[32] R. JIMÉNEZ-MORENO, R. A. CASTILLO. 深度学习语音识别在住宅助理机器人中的应用。*《国际人工智能杂志》*, 2022, 12(2): 585-592. http://doi.org/10.11591/ijai.v12.i2.pp585-592

Word count (excluding references): 4,862 words.

Peer-review record:

- *Fast-track status:* Not fast-tracked
- *First-round reviews received:* 3 reports
- *Revision cycles completed:* 3 rounds
- *Final version submitted:* July 28, 2025