



# The Design of Gesture Recognition-Based Mecanum-wheeled Vehicle

Zhouhan Tang

School of Information Science and Engineering, Harbin Institute of Technology, Weihai,  
264209 Weihai City, Shandong Province, China  
2023211386@stu.hit.edu.cn

**Abstract.** In the context of the rapid development of intelligent robot interaction technology, traditional remote-control methods have problems such as complex operation and poor interactivity. This study proposes a gesture-controlled Mecanum wheel car design based on the Mediapipe framework. Mediapipe is used to capture the key points of the hand in real time, and gesture recognition is performed by comparing the coordinates of the key points to achieve six command controls such as forward, backward, and in-place turning. A computer camera is used to capture images and perform gesture recognition, a Wi-Fi module is used to communicate with the car, and an ESP32 microcontroller is used to build the car's main control system and motor drive circuit. The test results show that the system can accurately recognize gesture commands within the same Wi-Fi range, the car responds quickly and can flexibly complete omnidirectional movement. The research results provide an intuitive and natural solution for human-computer interaction of intelligent cars and have application potential in scenarios such as robot control and educational demonstrations.

**Keywords:** Mecanum-wheeled car; gesture recognition; Mediapipe; contactless control.

## 1 Introduction

The concept of human-computer interaction has been around since the birth of computers, and its development trajectory has closely followed the iterative evolution of computer technology. From the early punched paper tape to the command line interaction, to the graphical user interface interaction that occupies the mainstream position today, human-computer interaction has the significant characteristics of direct control and what you see is what you get. In the traditional interaction mode, computers are in a core dominant position, and people mainly use devices such as buttons, mice and keyboards to complete interactive operations. However, these interaction methods have certain usage thresholds and require humans to adapt to the machine operation logic, which is essentially different from the natural communication methods that humans are accustomed to [1]. Thanks to the development of human-computer interaction technology and the promotion of the "people-centered" concept, more and more products have

begun to start from user needs and incorporate operating comfort into core design considerations, aiming to create an interactive experience that is more in line with human behavior habits.

As an important and basic form of non-verbal communication, gestures have quickly become a research hotspot in the field of human-computer interaction due to their natural intuitiveness and naturalness [2]. By capturing the motion information of hands and fingers through digital image processing technology or sensing the actual motion trajectory with the help of sensors, gestures are converted into machine-recognizable operation instructions. Gesture interaction breaks through the limitations of traditional interaction methods and makes the interaction between people and machines return to nature. In order to give full play to the application potential of gestures in human-computer interaction, it is necessary to extract gesture features and analyze the interaction logic and use gesture recognition technology to build an accurate and efficient control system.

Currently, gesture recognition technology is divided into traditional digital image processing methods and deep learning methods. Gesture recognition technology based on traditional methods relies on image segmentation, region selection, morphological processing, etc., and uses convexity defects for gesture recognition. It has low computing power requirements, but the feature extraction effect is poor when there is a lot of background noise in the image [3]. The gesture recognition method based on deep learning uses convolutional neural networks for feature extraction, which has high accuracy and strong robustness. It can adapt to various background environments and has achieved good results in the field of gesture recognition. Therefore, methods based on deep learning are widely used [4].

The birth of the Mecanum wheel is similar to the development trend of human-computer interaction. The Mecanum wheel is an omnidirectional mobile wheel with multiple sets of centrally symmetrical oblique rollers installed on its rim. By controlling the force direction of the oblique rollers on different wheels, the resultant force vectors of the four wheels can be freely superimposed on the plane to synthesize the moving force in any direction. Therefore, the car equipped with the Mecanum wheel can achieve forward and backward, left and right, oblique movement and rotation in place, and has three degrees of freedom in the plane [5]. Compared with the differential drive or tracked structure used in traditional cars, the omnidirectional mobility of the Mecanum wheel realizes the transformation from "mechanically constrained movement" to "natural and flexible movement", which is similar to the evolution of human-computer interaction from traditional device operation to gesture interaction, and both reflect the "people-centered" design concept.

Based on this, this paper proposes a design of a Mecanum wheel car based on gesture recognition control. When the Mecanum wheel car is controlled by gesture recognition technology, the user can achieve omnidirectional movement and precise positioning of the car with simple gesture commands, which not only significantly improves the convenience and intuitiveness of control but also extends the naturalness of human-computer interaction to the field of object operation. It is a practice of the "people-centered" design concept in the field of robot control.

The research is mainly divided into two parts: the design of the Mecanum wheel car and the implementation of gesture recognition. In terms of gesture control, the simplest and most intuitive gesture is 012345, as shown in Figure 1. Considering the left and right hands, a total of 36 different commands can be issued. In the design, six independent gestures are selected, namely 1, 2, 3 for the left hand and 1, 2, 3 for the right hand, corresponding to moving forward, translating to the left, rotating clockwise, moving backward, moving right, and rotating counterclockwise, respectively. The remaining idle gesture combinations can add commands according to actual conditions.

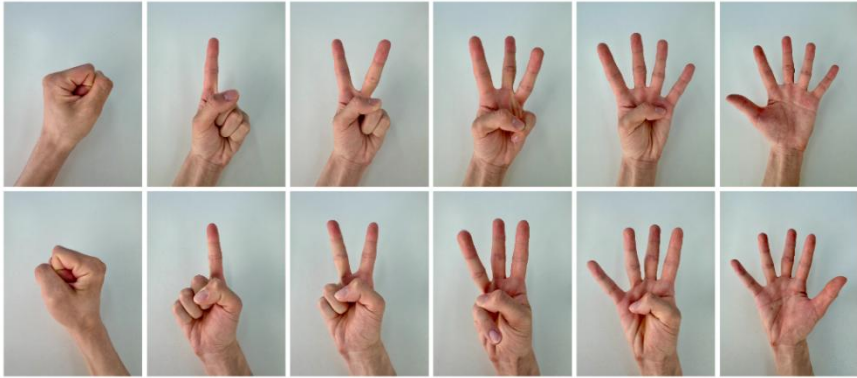


Fig. 1. Left and right hand 0-5 gestures.

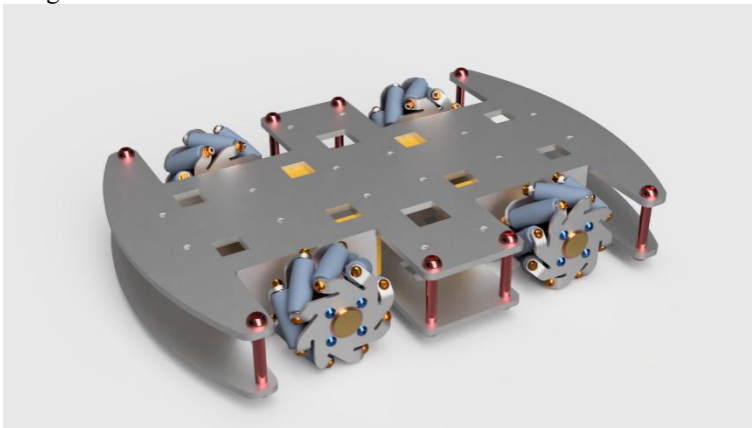
## 2 Mecanum-wheeled Vehicle Design

This article introduces the design idea of the Mecanum wheel car. The car uses four Mecanum wheels and is equipped with four motors of the same model. The four Mecanum wheels adopt a symmetrical X-shaped layout. The four wheels are distributed in a rectangular shape. The rollers are tilted  $45^\circ$ , and the tilting direction is symmetrical [6]. Relative to the center of the car, the roller of the left front wheel tilts outward (from upper left to lower right), the roller of the right front wheel tilts inward (from upper right to lower left), the roller of the left rear wheel tilts inward (from lower left to upper right), and the roller of the right rear wheel tilts outward (from lower right to upper left). The ESP32 motherboard and Wi-Fi module are installed on the car. The Wi-Fi module receives the gesture recognition instructions from the computer and controls the motor drive through the ESP32 motherboard. At the same time, a lithium battery box is installed to power the motherboard.

The Mecanum wheel consists of several key components, including two side support frames, a central shaft, an oblique roller, long and short single-head bolts, and anti-loosening nuts [7]. The central shaft connects the two side support frames and the motor drive shaft. Considering the metal strength, it is made of medium carbon steel. The two support frames are used to fix eight sets of oblique rollers. Since sheet metal processing is required, they are made of aluminum. The long single-head bolt is used to fix the two

support frames. Only one end is processed with threads for connecting nuts. The other end is designed with a hexagonal head for easy tightening of tools. The middle screw part is smooth and has no threads, which plays a supporting and positioning role. The short single-head bolt is used as the roller of the roller. Considering the need for wear resistance, stainless steel is used as the material. Since the Mecanum wheel car must vibrate frequently during the driving process, the nut of the fixing bolt needs to use an anti-loosening nut. A nylon ring is embedded at one end of the inner hole of the nut. The inner diameter of the nylon ring is smaller than the outer diameter of the thread. When tightened, the nylon ring is squeezed and deformed, and a radial clamping force is generated on the bolt thread, which can prevent the Mecanum wheel structure from loosening due to vibration. The roller is the core component of the Mecanum wheel drive, which needs to balance elasticity and friction coefficient. Polyurethane is selected as the material, which has excellent elasticity and wear resistance. Although the strength is relatively low, the central stainless-steel screw improves the overall stiffness to meet the load requirements.

The frame is made of two layers of aluminum plates. The lower plate fixes four motors and a transmission shaft. A motor groove is set on the plate. One end of the transmission shaft is connected to the motor, and the other end is connected to the central axis of the Mecanum wheel. The outer dimensions of the upper plate are the same as those of the lower plate. There are multiple threaded through holes designed on the plate to fix the embedded motherboard and battery box. Both plates have multiple openings to facilitate the motor wires to pass through the plate and connect to the motherboard. The overall structure adopts a symmetrical design to weaken the directionality. Each of the four Mecanum wheels can be independently controlled by the motor. Both in appearance and function, it meets the characteristics of omnidirectional movement. The four sides of the frame adopt arc profiles and protrude from the wheels to prevent the Mecanum wheel from being damaged by impact. The upper and lower plates are connected and fixed with 8 pairs of bolts and nuts, which are distributed around the frame for fastening. The modeling was carried out using Autodesk Fusion software, as shown in Figure 2.



**Fig. 2.** Mecanum-wheeled car

### 3 Gesture Recognition Algorithm Design

In terms of software, in order to control the car through gestures, it is necessary to implement the writing of the following two module codes: positioning and tracking the position of human hands and performing gesture recognition and controlling the direction and speed of the four motors based on the results of gesture recognition.

This study focuses on the gesture recognition part. Regarding gesture recognition, it is divided into five steps: real-time image capture, hand detection, left and right-hand judgment, outstretched finger index detection, and image display, as shown in Figure 3. Considering the recognition speed and accuracy, this paper adopts the hand detection solution based on OpenCV provided by Mediapipe to detect 21 hand key points, numbered 0 to 20, as shown in Figure 4. First, process the image transmitted by the camera and flip the image horizontally to make the video more in line with the habit of selfie. Then get the height ( $h$ ), width ( $w$ ) and number of channels ( $c$ ) of the image for subsequent coordinate conversion. Since OpenCV uses the BGR color space by default and Mediapipe uses RGB, the image must be converted from BGR to RGB [8]. Then detect whether there is a hand in the image. If at least one hand is detected, enter the loop processing. At this time, read the hand key points recognized by Mediapipe, and draw the key points and hand skeleton on the original image. Traverse each detected hand, access the key point data synchronously through the subscript, and obtain the confidence score of the hand being the left/right hand. Only when the confidence is greater than or equal to 0.8 will the processing continue. Traverse all the key points of the hand and convert the relative coordinates  $[0, 1]$  to the actual pixel coordinates  $(x, y)$  in the image. Then count the number of fingers of the specified hand that are open. As shown in Figure 4, the IDs of the key points at the end of the five fingers are: thumb: 4; index finger: 8; middle finger: 12; ring finger: 16; little finger: 20. For each finger, the relative position of the fingertip (tip  $i$ ) ( $i=4, 8, 12, 16, 20$ ) and the node (tip  $i-2$ ) one joint below the fingertip can be used to determine whether the finger is straight. For the thumb, the thumb is open horizontally, and the  $x$  coordinate is used to judge that when the left hand is open, the  $x$  coordinate of the fingertip is greater than the coordinate of the root of the finger, and the right hand is the opposite. For the index finger to the little finger, they are open vertically, and the  $y$  coordinate is used to judge that if the  $y$  coordinate of the fingertip is less than the root (the origin of the image is in the upper left corner,  $x$  is positive to the right, and  $y$  is positive downward), it means that the finger is open. Compare each finger one by one and finally accumulate the number of fingers extended as the operation gesture. The gesture numbers are drawn on the computer screen and transmitted to the car mainboard through the WiFi module. In order to avoid the car's motion response lag caused by different gestures at adjacent moments, it is set to send gesture control commands to the car every 2 seconds.

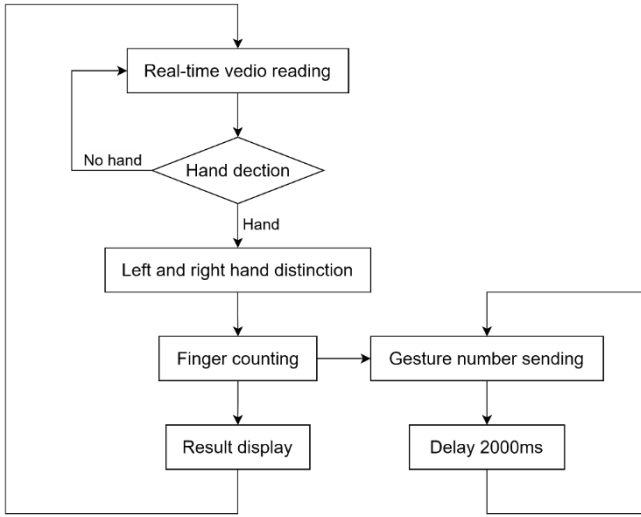


Fig. 3. Gesture recognition and transmission process

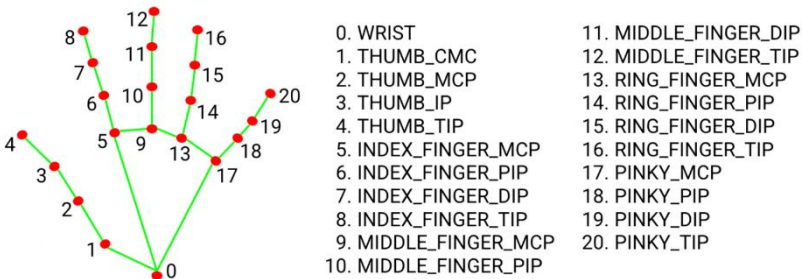


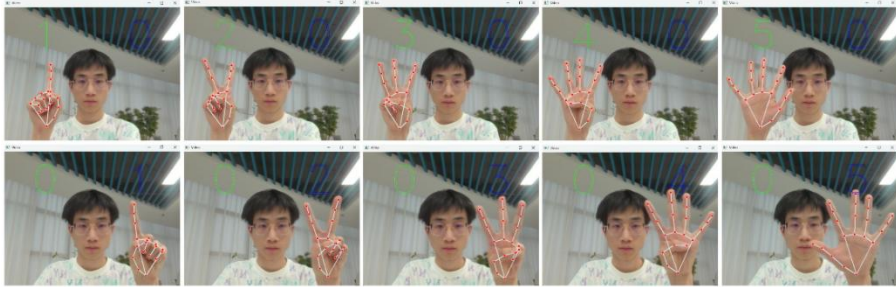
Fig. 4. 21 key points of the hand provided by Mediapipe

Regarding motor control, it is necessary to set up a combined action module for four Mecanum wheels. When moving forward and backward, all wheels rotate in the same direction at the same speed. The oblique force generated by each wheel can be decomposed into forward and backward components and left and right components. Because the layout of the four wheels is symmetrical, the left and right components cancel each other out, leaving only the forward or backward components to be superimposed, thereby achieving forward or backward movement. When translating to the left, the left front wheel turns forward, the right front wheel turns backward, the left rear wheel turns backward, and the right rear wheel turns forward. Under this rotation combination, the left component of the oblique force generated by all wheels will be superimposed, while the forward and backward components will cancel each other out. The effect of the superposition of all wheel forces is a left-facing resultant force. Similarly, when translating to the right, the left front wheel turns backward, the right front wheel turns forward, the left rear wheel turns forward, and the right rear wheel turns backward. When

rotating clockwise in place, the left front wheel and the left rear wheel turn backward, and the right front wheel and the rear wheel turn forward. When rotating counterclockwise in place, the left front wheel and the left rear wheel turn forward, and the right front wheel and the right rear wheel turn backward [9]. When the wheels rotate in this combination, a torque around the center of the vehicle is generated, driving the car to rotate in place. When moving diagonally, by adjusting the speed and direction of each wheel, the resultant force in any direction can be synthesized, thereby achieving straight-line movement at any angle [10].

## 4 Experiments and Results

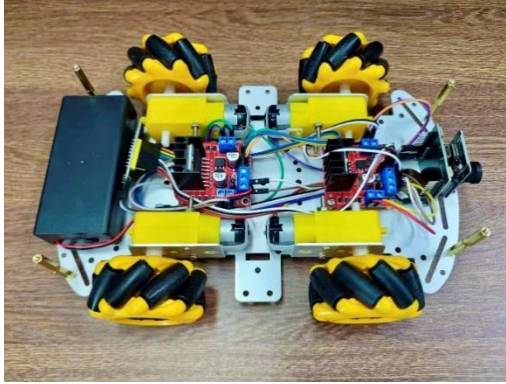
As shown in Figure 5, the processing results of left hand gestures 1-5 and right hand gestures 1-5 are presented. The red dots in the figure are the key points of the hand obtained by image processing, the white lines are the skeletons connecting the key points, the green numbers on the upper left are the left hand gesture numbers, and the blue numbers on the upper right are the right hand gesture numbers.



**Fig. 5.** The recognition results of left and right hand gestures 1-5

After repeated tests, the accuracy of gesture recognition is relatively high, and the program can provide real-time feedback. Even when the hand image and the facial image partially overlap and the skin color is almost the same, the program can still correctly recognize the gesture. It has strong robustness and precise feature distinction capabilities and is reliable in real complex environments.

The assembled car is shown in Figure 6. When the car and the computer are connected to the same Wi-Fi network environment, the left-hand gesture 123 and the right-hand gesture 123 are input in sequence. Observation indicates that the car can accurately execute the preset action instructions, including moving forward, translating to the left, rotating clockwise, moving backward, moving right, and rotating counterclockwise. The system is set to transmit instructions every 2 seconds. During the test, when the gesture signal is sent, the car responds quickly, with a delay of no more than 200 milliseconds, and the action execution has a high real-time performance.



**Fig. 6.** The assembled Mecanum-wheeled car

## 5 Limitations and Future Outlooks

In the study of gesture recognition, this study uses the method of key point coordinate comparison to determine the finger extension state. Experimental results show that this method can accurately recognize natural upward or oblique upward gestures, but when the gesture appears in the image frame with a downward posture, it will cause judgment errors because the coordinate system is difficult to effectively distinguish the spatial front and back relationship. Based on this, subsequent research intends to introduce a vector representation method, and the finger end joint point (tip  $i$ ) ( $i=4, 8, 12, 16, 20$ ) and the node (tip  $i-2$ ) separated by one joint constitute the finger feature vector. At the same time, a reference vector is added. The reference vector is composed of the node at the base of the palm (tip 0) and the node (tip  $i-3$ ) separated by two joints from the fingertip. By calculating the angle between the two vectors, the finger extension state can be more accurately determined.

In addition, during the gesture recognition system testing phase, experiments on image occlusion scenes showed that although the Mediapipe model can achieve accurate segmentation of the hand and facial areas, when the hand signs of both hands overlap in a large area in front and back, the system can only complete the detection and positioning of a single hand [11]. This recognition result deviates from the preset requirement of simultaneous perception of both hands. Since a gesture instruction set based on two-handed collaboration was not constructed in this study, the current recognition error does not affect the core research objectives. However, this phenomenon has potential constraints on the construction of expanded multi-gesture instructions. In the future, multiple cameras or depth cameras installed at other angles can be introduced to obtain three-dimensional information of the scene, and a three-dimensional model of the hand can be constructed through a stereoscopic vision algorithm. In this way, even if the hands overlap on the two-dimensional plane, the position and posture of the hands can be distinguished from the three-dimensional space, thereby achieving accurate recognition of the hands under front and back occlusion.

## 6 Conclusion

This study successfully designed and implemented a gesture recognition remote-controlled Mecanum wheel car system based on Mediapipe. Through real-time capture and analysis of human hand joints by Mediapipe, gestures are converted into control commands, and the Mecanum wheel car is accurately driven to complete forward, backward, left and right translation, and rotation in situ. The experimental results show that the system can effectively recognize left and right-hand gestures 123, and the car responds quickly, which effectively verifies the feasibility of combining gestures with Mecanum wheels for human-computer interaction control.

This study also expands the application scenarios of Mediapipe in the field of intelligent control and provides a new engineering implementation case for human-computer interaction technology based on computer vision. The system simplifies the traditional remote control operation process and realizes intuitive and convenient non-contact control. It can be applied to robot control, education and teaching and other scenarios.

However, this study still has some limitations. The gesture recognition process requires the gesture to be in a natural upward state, which has certain posture constraints; and the system relies on the Mediapipe pre-trained model for two-dimensional image processing. Without adding additional input information, it is difficult to effectively recognize the gesture features of the scene where both hands overlap front and back; in addition, the current system only implements basic one-hand gesture control, and the recognition and processing of complex combination commands has not yet been perfected.

To address the limitations mentioned above, future research can be carried out from three aspects: first, optimize the gesture recognition algorithm and use the directionality of the vector to solve the posture constraints on the hands; second, introduce multi-modal sensor fusion technology, such as combining depth cameras to collect depth information, and enhance the system's judgment of overlapping hand scenes; third, expand the gesture control instruction library to achieve more complex and diversified gesture operations, and promote the construction of a more natural and convenient gesture control system.

## References

1. A. Wang and J. Ye, Human-Computer Natural Interaction Design Practice Based on Unconscious Design Concept, in M. Kurosu (Ed.), Human-Computer Interaction. Theory, Methods and Tools, Lecture Notes in Computer Science, 12762, Springer, Cham, (2021). doi:10.1007/978-3-030-78462-1\_1.
2. L. Guo, Z. Lu, and L. Yao, "Human-machine interaction sensing technology based on hand gesture recognition: A review," *IEEE Trans. Hum. -Mach. Syst.* 51(4), 300-209 (2021)
3. R. Tripathi and B. Verma, "Survey on vision-based dynamic hand gesture recognition," *Visual Comput.* 40(6), 6171–6199(2024). doi:10.1007/s00371-023-03160-x.

4. M. Gil-Martín et al., "Hand Gesture Recognition Using MediaPipe Landmarks and Deep Learning Networks," in Proc. 17th Int. Conf. Agents and Artificial Intelligence (ICAART), 3, (2025)
5. J. D. Morin, Control of Mecanum Wheels in Arbitrary Position, Angle and Number for Complete Omni-Directional Motion, Ph.D. dissertation, Massachusetts Institute of Technology, 2024.
6. Y. Shi, J. Li, Y. Yang, and Y. Han. Design of a Solar-Powered Car Based on Gesture Recognition. In 2024 International Conference on Image Processing, Computer Vision and Machine Learning (ICICML), Shenzhen, China, 2024, pp. 1843-1847. DOI: 10.1109/ICICML63543.2024.10957989.
7. T. Giurgiu et al., "Mecanum wheeled platforms for special applications," Int. Conf. Knowledge-Based Organization, 28(3), (2022)
8. A. Bhawarathi et al., "Hand Gesture Controlled Wheelchair Using Image Processing," 2024 IEEE Symposium on Industrial Electronics & Applications (ISIEA), Kuala Lumpur, Malaysia, 2024, pp. 1-6, doi: 10.1109/ISIEA61920.2024.10607210.
9. M. Y. Naing, A. S. Oo, I. Nilkhamhang, and T. Than. Development of Computer Vision-Based Movement Controlling in Mecanum Wheel Robotic Car. In 2019 First International Symposium on Instrumentation, Control, Artificial Intelligence, and Robotics (ICA-SYMP), Bangkok, Thailand, 2019, pp. 45–48. DOI: 10.1109/ICA-SYMP.2019.8646254.
10. T. N. van der Spijk, Model-Based-Control for Trajectory Tracking with a Mecanum Wheeled Vehicle: A performance comparison between kinematic and dynamic model-based control, M.S. thesis, Embedded Systems, Delft University of Technology, Netherlands, 2024.
11. S. Uke, A. Shaikh, H. Rayate, A. Kamble, and S. Rahane, "Towards Touchless Interaction: Implementing Hand Gesture Recognition for Presentation and Media Control," in Proc. Int. Conf. Emerging Smart Computing and Informatics (ESCI), Pune, India, 1-6 (2025). doi:10.1109/ESCI63694.2025.10988099

**Open Access** This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

