

MODEL AND METHOD OF INCLUSIVE INTERACTION WITH VISUAL AND TEXTUAL CONTENT BASED ON CONTEXTUAL GESTURE INTERPRETATION

Abstract. Background. Contactless gesture control, as an auxiliary channel of interaction with an interface, has now become widely used in various areas of life, as it provides a convenient, safe, and, if necessary, inclusive way of interacting with visual, graphical, and textual content. **The subject of the study** is the process of contactless control of basic navigation scenarios for working with visual, textual, and web content using hand tracking methods and contextual interpretation of gesture events. **The objective** of the study is to develop and experimentally validate a method of contactless interaction with visual and textual content based on RGB hand tracking and contextual interpretation of gestures for implementing practical user scenarios. The main research task of the study is to develop a conceptual model of context-adaptive inclusive user interaction with an information system and the external environment. **Methods.** Functional modeling, computer vision methods, analysis of hand keypoints, geometric rules for gesture recognition, and experimental testing were used. **Results.** A method is proposed in which a gesture event is converted into a command only after determining the type and analyzing the context, namely the technical and scenario context. The experimental validation was carried out using five basic gestures and three scenarios. **Conclusions.** The obtained results confirmed the operability of the method as an auxiliary channel of contactless interaction with visual and textual content.

Keywords: contactless interaction, tracking, gesture recognition, human-computer interaction, Media Pipe Hands, gesture control, interface, model, method.

1. Introduction

1.1 Relevance of the topic

The digitization of education, public services, everyday environments, and professional activities is leading to an increase in the number of scenarios in which users interact with visual and textual content via electronic devices (Human-Computer Interaction, HCI). At the same time, there is a growing need not only to create digital content but also to develop convenient, inclusive, and secure ways to interact with it, with so-called contactless technologies being in particularly high demand.

According to estimates by the World Health

Organization, about 1.3 billion people – or approximately 16% of the world's population – have significant disabilities [1]. For some of these users, traditional input methods may be inconvenient, inaccessible, or physically limited. In this context, gesture-based contactless interaction can be used as an auxiliary method of interface control, especially when a user is able to perform a limited set of hand movements but cannot effectively use a mouse, keyboard, or touchpad. In addition, gesture-based interfaces hold promise for rehabilitation, training, and adaptive environments, where hand movement can simultaneously serve as both a command input and a means of assessing motor activity (Table 1).

Table 1 – An Overview of the Applications of Contactless Human-Machine Interaction Interfaces

Field of application	Situations in which contactless methods are most in demand
Health Care and Surgical Environments	sterile viewing of CT/MRI/X-ray images, interface control without a mouse or keyboard, working while wearing gloves
Rehabilitation	fine motor skills training, hand movement assessment, gamified exercises, recovery from injuries or neurological disorders
Education and Inclusive Learning	lectures, presentations, interactive whiteboards, LMS, maintaining student attention, learning for students with special educational needs
Public Information Displays and Kiosks	public Service Centers, hospitals, universities, museums, exhibitions, transportation hubs, self-service terminals
Everyday and smart-home scenarios	controlling media, screens, and household devices when your hands are full or you don't want to use your voice
Automotive Industry	control volume, calls, and navigation without having to search for elements on the touchscreen
Industry, Manufacturing, and Robotics	operating robots, manipulators, and equipment in hazardous, contaminated, or hard-to-reach environments
VR/AR and Multimedia Environments	interaction with virtual objects, educational simulators, professional XR training simulators, controller-free interaction
Special Needs and Assistive Technologies	assistive computer control, communication aids for people with speech or hearing impairments, adaptive workstations, speech therapy or motor skills training devices

It is worth noting the relevance of this issue for educational and presentation environments. A prime example of its use is digital learning materials, presentations, video lectures, interactive whiteboards, distance learning systems, and public information screens, which are becoming widespread tools in the educational process. In such situations, teachers or users often need to navigate the material without constantly having to use a keyboard, mouse, or touchscreen – to switch slides, scroll through text, open interface elements, confirm selections, or zoom in or out on a visual object.

Therefore, contactless interaction with content can be viewed as an additional control channel that enhances the flexibility and ergonomics of working with educational and informational materials.

A separate group of scenarios relates to public and healthcare settings, where reducing direct physical contact with shared surfaces is crucial. In healthcare facilities, the issue of touching shared surfaces is particularly sensitive because, according to the WHO, among every 100 patients in emergency departments, at least 7 patients in high-income countries and 15 patients in low- and middle-income countries may contract at least one healthcare-associated infection [2]. At the same time, proper hand hygiene can prevent up to 50% of infections that can be avoided during the provision of medical care [3]. Although contactless interfaces do not replace hygiene measures, they can reduce contact with shared surfaces and increase the acceptance of digital services in environments with heightened hygiene requirements.

The limitations of existing solutions include the sensitivity and hardware characteristics of the recording devices used for subsequent hand tracking. The quality of gesture recognition depends on lighting, background, video stream resolution, distance from the camera, hand angle, partial finger occlusion, and movement speed. Review studies on hand gesture recognition show that, despite advances in computer vision and deep learning, challenges related to robustness to environmental changes, gesture complexity, computational efficiency,

and adaptation to different users remain unresolved. Therefore, for practical interaction with content, it is not enough to simply identify which gesture the user is performing. It is also necessary to determine exactly how this gesture should be interpreted in the system's current operating mode and within a specific interaction scenario.

1.2 Literature Review and Problem Definition

This subsection analyzes literature sources and recent studies devoted to hand gesture tracking approaches that differ in the type of input data, hand position estimation accuracy, hardware cost, ease of use, and suitability for large-scale application. The analyzed approaches include:

- tracking using RGB cameras;
- tracking using depth sensors;
- tracking using wearable sensors.

The analysis made it possible to identify several local unresolved issues, which can be systematized as follows:

- RGB-based tracking is accessible and easy to implement, but it depends on illumination, hand orientation, distance to the camera, and video stream resolution;
- most studies treat the recognized gesture as the final result, rather than as an intermediate gesture event that must be further transformed into a control command;
- in applied systems, gestures are often tied to one narrow scenario and a fixed set of commands;
- insufficient attention is paid to the fact that the same gesture may have different meanings depending on the active system mode, content type, hand role, and previous interaction state;
- complete scenarios of interaction with visual and textual content, such as file operations, web navigation, and drag-and-drop, remain insufficiently studied;
- the influence of external recording conditions on gesture recognition accuracy and the stability of applied interaction scenarios requires separate experimental verification (Fig.1).

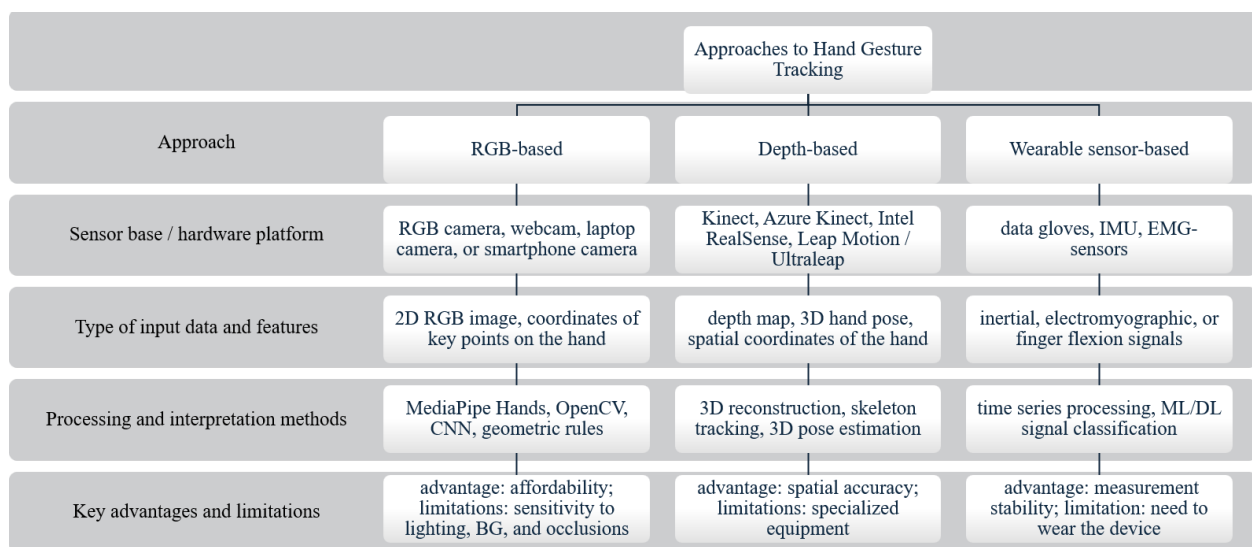


Fig. 1. Approaches to contactless hand gesture tracking

This makes it possible to formulate the general problem in the field of contactless gesture-based content control in human-machine interaction: the system should not only recognize a gesture, but also determine which interface action this gesture corresponds to in the current operating mode, for a specific type of content, and within a particular interaction scenario. This unresolved problem determines the aim of the present study.

In [4], it is shown that depth sensors, in particular Leap Motion, provide sufficiently accurate estimation of the spatial position of the hand and fingers based on three-dimensional hand geometry. However, this approach requires specialized and relatively expensive equipment, which is more difficult to integrate into typical user environments. Therefore, such an approach is less suitable for large-scale implementation in educational, public, or everyday interfaces.

In [5], an approach based on wearable inertial sensors is considered. Its strengths include lower dependence on illumination, background, and hand occlusion in the camera frame. At the same time, the user has to wear an additional device, which complicates the use of such systems in public or short-term interaction scenarios. For this reason, wearable sensors are not always convenient for fast contactless interaction with information content.

In contrast to these approaches, works [6, 7] demonstrate the feasibility of using a standard RGB camera and methods for detecting hand keypoints. In particular, MediaPipe Hands makes it possible to obtain the coordinates of 21 hand keypoints without additional sensors. This approach is inferior to depth sensors in terms of spatial accuracy, but it has important advantages: accessibility, low cost, ease of integration, and the possibility of use on a typical computer or laptop. Therefore, RGB-based hand tracking is appropriate for developing an accessible method of contactless interaction with visual and textual content.

At the same time, even within the RGB-based approach, a number of unresolved issues remain. In [8] and [9], real-time gesture recognition solutions are proposed using MediaPipe, YOLOv11, Inception-v3, and LSTM. The main focus of these studies is the recognition of a gesture as a class, while the further transformation of the recognized gesture into an interface action is considered insufficiently. This is because these studies are mainly aimed at improving classification accuracy, rather than developing a complete cycle of user interaction with digital content.

The practical significance of contactless gesture-based control methods should also be emphasized. The literature describes a wide range of application areas, including the use of gesture analysis in educational environments to recognize teacher signals and support students' attention [10]; interaction with public web displays, where gestures are used for contactless navigation through information content [11]; inclusive offline learning platforms for hands-free interaction with educational materials [12]; rehabilitation and assistive systems for tracking hand movements during rehabilitation exercises [13] and controlling rehabilitation exergame systems [14]; assistive interaction for users with

disabilities in smart-home environments [15]; control of household or IoT devices using local or edge-based gesture processing, command execution, and feedback generation [16, 17]; and human-robot interaction, where gesture recognition is required to form a navigation action or an operation in the environment [18].

Thus, the analysis of literature sources shows that gesture-based approaches are applied in education, public information interfaces, inclusive learning platforms, rehabilitation, smart-home systems, IoT systems, and human-robot interaction. However, most of these solutions are focused on a narrow application scenario and a fixed set of commands.

The weak point of the analyzed studies is insufficient attention to the problem of contextual gesture interpretation. In most cases, the gesture is either recognized as a separate class or directly linked to a fixed command. However, in interaction with visual and textual content, the same gesture may correspond to different actions depending on the current system mode, content type, hand role, previous interaction state, and user scenario. Therefore, the unresolved problem is the development of a method that combines RGB-based hand tracking with contextual interpretation of gesture events for contactless interaction with visual and textual content. This problem defines the aim of the further research.

1.3 Aim and Objectives of the Study

The aim of the study is to develop and experimentally validate a method of contactless interaction with visual and text content based on RGB hand tracking and contextual interpretation of gestures for implementing practical user scenarios.

To achieve the stated goal, the following tasks must be addressed:

- develop a conceptual model of context-adaptive, inclusive user interaction with an information system and the external environment;
- develop a structural model of a contactless interaction method that encompasses capturing an RGB video stream, determining key points of the hand, calculating geometric features, recognizing gestures, contextually interpreting gesture events, and generating interface control actions;
- experimentally evaluate the impact of external conditions on gesture recognition accuracy, specifically the distance to the camera, lighting level, hand angle, and video stream resolution;
- analysis of the obtained results.

2. Materials and Research Methods

The object of research is the process of contactless management of basic navigation scenarios, related to work with visual, textual and web content by means of one-handed, multi-handed or scripted gesture control based on RGB tracking of the hand and interpretation of gesture events depending on the context and scenario.

The input data for the system is hand movements recorded by a regular RGB camera, and the output data consists of high-level control actions: hovering, selecting an element, moving an element, scrolling the page, confirming, canceling, or executing a script command.

The main hypothesis of the study is that the implementation of two levels of contextual interpretation of gestural events, which determine the states and conditions of receiving input data, will increase the practical suitability of contactless interaction with visual and textual content.

The selected context levels include the following:

- technical context of gesture recognition (illumination, camera distance, hand angle, video flow resolution, occlusion, background complexity, etc.);
- scenic context of gesture interpretation (system state - active or passive; type of operating mode - one-handed, two-handed, pointer mode, sequence mode, calibration; the class of the involved hand (dominant or non-dominant); preliminary data on the state of the system; the selected type of content for interaction is text, web, graphic, OS element; admissibility of action in the current state of the interface).

In the development and experimental validation of the method, the following assumptions and simplifications are adopted.

Condition 1. The interaction is carried out using a conventional RGB camera without the use of depth sensors, wearable devices, instrumented gloves or additional markers on the user's hand.

Condition 2. The user's hand is in the camera's field of view for unimpeded detection and presentation as a set of key points of the hand and geometric relationships between them.

Condition 3. A gestural event is not seen as a completed command. Only after analyzing the context, the gestural event is converted into a control action.

Condition 4. In the work, experimental evaluation is limited to a fixed set of gestures (OPEN_PALM, PINCH, FIST, PINCH_MIDDLE, THUMBS_UP) and user scenarios - basic system and navigation actions, work with text or web content, management of graphic objects.

The principal methods employed in the present study include functional modeling (applied in the construction of the models), computer vision techniques, geometric analysis of hand skeletal keypoints, and experimental validation.

For the acquisition of the input video stream, an RGB camera was utilized as a cost-effective and scalable solution for implementing contactless human-computer interaction (HCI). As substantiated in prior investigations [9, 19], gesture-based interaction with computing systems in educational, clinical, domestic, mobile, and assistive settings represents a rational approach, since it obviates the need for dedicated wearable hardware or complex sensory infrastructures.

The retrieval of 21 keypoint coordinates for the comprehensive description of the palm and phalanges per frame is accomplished via computer vision algorithms, thereby enabling real-time performance even on mobile computational platforms.

The viability of the MediaPipe Hands framework and the landmark-based methodology is corroborated by the findings reported in [14, 20].

The adoption of the landmark-based paradigm is justified by its capacity to shift the analytical focus from

holistic image processing to a compact feature-based hand representation, consequently reducing the dimensionality of the input data through the isolation of a precisely defined region of interest [9].

The methodological choices delineated in [9, 14, 20] are deemed essential, as they ensure sensor-independent operation, facilitate straightforward system integration, and render the approach applicable across conventional educational, domestic, public, and assistive environments. Nevertheless, it has been emphasized in the literature [14, 16, 20] that the practical efficacy of such systems remains contingent upon a number of factors, including ambient conditions, computational resource constraints, illumination levels, camera-to-user distance, real-time processing robustness, and the ergonomic suitability of the chosen gesture vocabulary. These considerations have been duly addressed within the framework of the proposed system and are effectively compensated for by the proposed context-aware interpretation mechanism.

To verify the effectiveness of the proposed method, three groups of test scenarios have been developed, which allow for the evaluation of the proposed method not only in terms of the accuracy of recognizing individual gestures, but also in terms of its ability to perform complete user actions while interacting with the operating system, text, web, and graphical content (scenarios are listed in order of increasing complexity):

- basic system and navigation actions (testing the ability to control operating system elements without using a mouse or keyboard – pointing, moving the cursor with the leading hand, clicking, and opening folders using basic gestures – THUMBS_UP, PINCH, PINCH_MIDDLE);
- working with text or web content (testing interaction with information content – hovering over a page element, selecting, scrolling, and navigating to another part of the content using basic gestures – THUMBS_UP, PINCH, FIST);
- manipulating graphical objects (testing the stability of control when selecting, holding, and moving an object – hovering over the object, triggering the selection action, moving the object to the desired area, and completing the action using basic gestures – THUMBS_UP, PINCH).

The experimental study of the proposed method takes into account the conditions that affect the quality of RGB hand tracking and the stability of gesture event generation. Since the system uses a video stream from a standard RGB camera, its performance depends on:

- the visibility of the hand in the frame;
- lighting;
- background;
- the distance to the camera;
- the angle of the hand;
- the resolution of the video stream.

The main limitations include partial obstruction of the hand or fingers, the hand extending beyond the frame, varying stability of individual gestures, and the transmission of commands at the operating system event level without considering the internal state of a specific application.

External factors include lighting, background, distance, and the angle of the hand, while internal factors include threshold settings, hand detection parameters, system operating mode, camera resolution, and computer performance. Controlling these factors allows for an accurate assessment of the method's performance and helps distinguish errors in the algorithmic interpretation of gestures from errors caused by shooting conditions or the hardware and software environment.

The basic conditions for conducting the study were defined as a constant camera resolution, a consistent distance between the user and the camera, a neutral background, a frontal hand position, and a fixed set of gestures. A video stream resolution of 1280×720 was used as the baseline setting, as it provides a balance between image detail and processing speed.

In a controlled series of measurements, the user held their hand approximately 0.5 m from the camera, the illuminance was about 153 lx, and the hand was positioned primarily facing the camera head-on.

First, the recognition accuracy of individual gestures was determined under baseline conditions; thereafter, individual influencing factors – illuminance, distance from the camera, hand orientation (Fig. 2), and video stream resolution – were systematically varied.

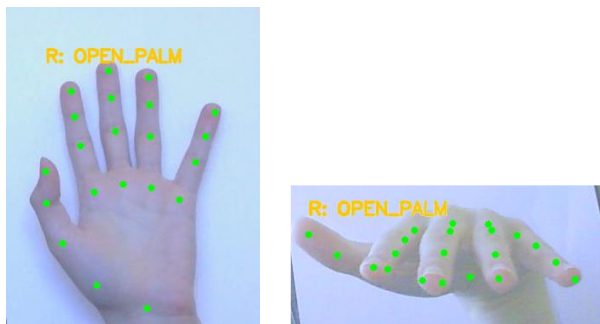


Fig. 2. An example of testing the method's capability when changing the perspective for the open-palm gesture

Next, the system's performance was tested in practical scenarios involving interaction with the operating system, as well as text, web, and graphical content. For each series, recognition accuracy, the number of omissions and errors, frame rate, response time, and scenario execution time were recorded. This methodology made it possible to compare the system's performance under standard and modified conditions, identify the most stable gestures, and determine the factors that most significantly reduce the accuracy and speed of the proposed method. Each scenario was run 10 times using gesture control and 10 times using traditional mouse and keyboard control. The table shows the average execution time, the total number of errors, and the number of critical errors.

3. Results of Development and Experimental Study of the Proposed Contactless Interaction Method

Within the scope of this study, contactless gesture interaction is considered not as an isolated method for recognizing individual hand positions, but as one of the channels for context-adaptive user interaction with an information system. The general logic of such interaction is presented as a conceptual model that takes into account the user's input request, individual needs, the application scenario, the user's context, the interpretation of input data, the selection of an assistance strategy, and the generation of feedback through a multimodal interface.

The proposed conceptual model (Fig. 3) demonstrates that effective interaction with an information system depends not only on the method of entering a command, but also on the context in which it is used. Such context may include the type of information content, the level of accessibility of the environment, the presence of obstacles, the current task scenario, the user's sensory or motor capabilities, as well as the available feedback channels.

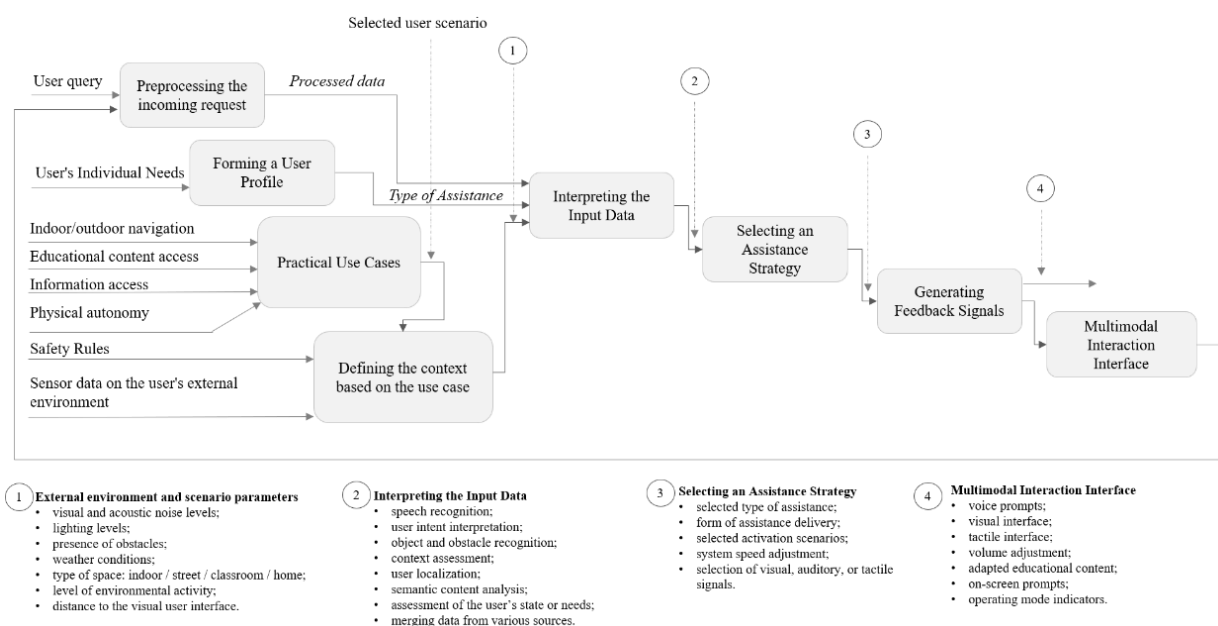


Fig. 3. A Conceptual Model of Context-Adaptive, Inclusive User Interaction with an Information System and the External Environment

In such a system, gesture control can serve as one of the input channels, complementing voice, visual, auditory, or tactile means of interaction.

Within the framework of the conceptual model presented, this article focuses on one of its functional components – the contactless input channel, in which the user's hand movements are converted into commands for interacting with visual and text content. Unlike the general model of inclusive interaction, where various types of sensory data and feedback signals can be used, this study examines RGB hand tracking, the generation of basic gesture events, and their contextual interpretation as interface control commands.

Thus, the proposed method specifies the block responsible for interpreting input data and generating feedback control actions in a context-adaptive interaction system.

Its purpose is to convert hand movements captured by an RGB camera into high-level interface actions: pointing, selecting, scrolling, confirming, canceling, or executing a scripted command.

The model proposed in this paper for a method of non-contact interaction with visual and text content, based on RGB hand tracking and contextual

interpretation of gestures, is presented as a functional diagram consisting of five functional blocks:

- RGB frame acquisition / preprocessing;
- hand detection / keypoint identification;
- gesture event formation;
- mode- and context-based interpretation of the gesture event;
- generation of a remote-control command / OS action.

The model's primary function is to convert the user's movements, captured by an RGB camera, into interface control actions, taking into account the active operating mode, the role of the hand, the type of content, and the current interaction scenario.

A key distinction of the proposed model is the presence of a mode-context interpretation block, in which an elementary gesture event is mapped to a control action not directly, but taking into account the system's active state, operating mode, previous state, the presence of confirmation, content type, and interaction scenario.

The proposed IDEF0 model illustrates the sequence of steps for converting an RGB video stream into a high-level remote-control command for visual and text content (Fig. 4).

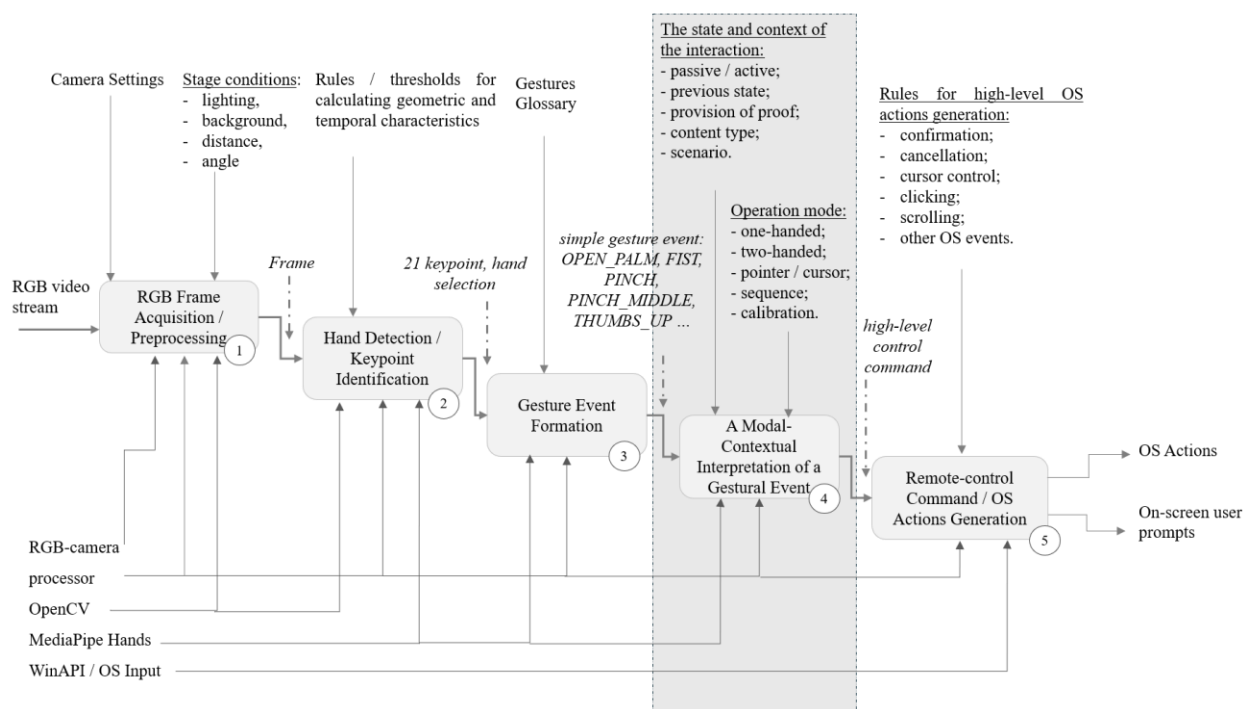


Fig. 4. A Generalized Model of a Non-Contact Interaction Method with Visual and Textual Content Based on RGB Hand Tracking and Contextual Gesture Interpretation

Block 1, “RGB Frame Acquisition / Preprocessing,” acquires a frame from an RGB camera and prepares it for further analysis. At this stage, the camera configuration and scene conditions – including lighting, background, distance to the user, and brush angle – are taken into account. The result of this block's operation is a prepared RGB frame, which is passed on to the brush detection stage.

Block 2, “Hand Detection / Keypoint Identification,” detects the hand in the frame and determines the coordinates of the hand's keypoints. The

output of this block is a compact representation of the hand in the form of a set of keypoints, and it also selects or identifies the hand for further processing. This stage is an intermediate step between image analysis and the generation of a gesture event.

Block 3, “Gesture Event Formation,” converts the coordinates of keypoints into a basic gesture event. This is done using rules and thresholds for calculating geometric and temporal features, as well as a dictionary of basic gestures. Such gesture events include, in particular, OPEN_PALM, FIST, PINCH,

PINCH_MIDDLE, THUMBS_UP, and other valid hand positions or combinations (Fig. 5). At this stage, the system does not yet generate an operating system command, but only determines which basic gesture event was detected.

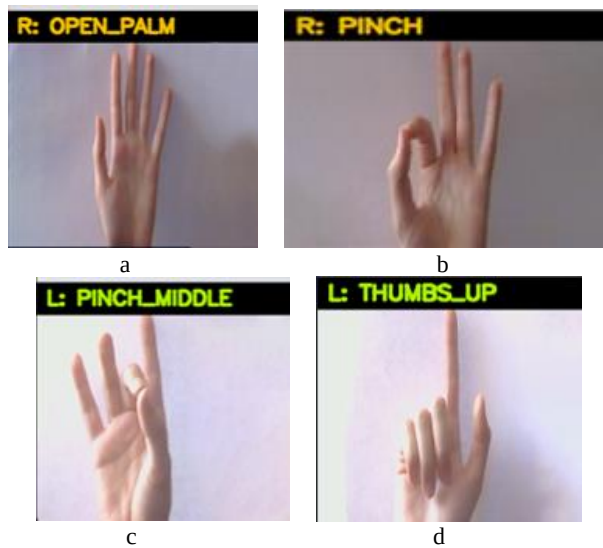


Fig. 5. Basic hand gestures: a – open palm (OPEN_PALM); b – pinch with the thumb and index finger (PINCH); c – pinch with the thumb and middle finger (PINCH_MIDDLE); d – thumbs up (THUMBS_UP)

A key element of the model is Block 4, “Mode-Contextual Interpretation of a Gesture Event.” It is in this block that a basic gesture event is matched to a possible control command – not directly, but taking into account the state and context of the interaction. These factors include the system’s passive or active state, the previous state, the presence of confirmation, the content type, and the current operating scenario. Additionally, the active operating mode is taken into account: one-handed, two-handed, pointer control mode, sequence input mode, or calibration mode.

Within block 4, the same basic gesture event can take on different functional meanings depending on the active mode and context. For example, a gesture can be

interpreted as confirmation, cancellation, a sequence element, activation of pointer mode, or part of a content control command. If the system is in a passive state or the confirmation conditions are not met, no control action is generated. This helps reduce the number of accidental activations and increase the predictability of the interaction.

Block 5, “Remote-control Command / OS Actions Generation,” converts a high-level control command – received after mode- and context-based interpretation – into a specific action of the operating system or interface. Such actions include switching operating modes, buffering sequences, confirmation, cancellation, cursor control, clicking, scrolling, and other operating system events. The model’s outputs are operating system actions and on-screen user prompts, which provide feedback during interaction.

The model’s control inputs include the camera configuration, scene operating conditions, feature calculation rules and thresholds, the gesture vocabulary, the state and context of the interaction, operating modes, and the rules for generating high-level actions of the interface or operating system. The execution mechanisms include an RGB camera, a computer, OpenCV, MediaPipe Hands, a geometric rules module, a contextual interpreter, and a mechanism for transmitting events via WinAPI/OS Input. The proposed model describes not only the technical process of gesture recognition but also the full cycle of remote, contactless content control. Its key feature is mode-contextual interpretation, thanks to which a gesture event is converted into a control action only after checking the active mode, previous state, confirmation, content type, and interaction scenario.

Within the proposed model, passive mode corresponds to a standby state in which the system analyzes the video stream but does not generate OS control events (Table 2). The transition to active mode occurs only after a valid gesture event or activation command is detected. This approach reduces the likelihood of false triggers, since random hand positions are not directly converted into interface actions.

Table 2 – Operating Modes of the Contactless Interaction System

Mode	Purpose	Typical Actions
Passive	Anticipation, frame analysis without forming commands	clearing temporary states, updating the OSD
Active (one-handed)	Executing simple commands with one hand	select, cancel, confirm, single gesture
Active (two-handed)	Assigning roles to each hand	dominant hand – positioning, non-dominant hand – confirm/click
Mouse/Pointer Mode	Cursor control	cursor movement, PINCH, PINCH_MIDDLE, FIST
Sequence Mode	Entering more complex commands	gesture accumulation, confirm, cancel
Calibration	Customizing thresholds for the user	updating threshold values

Active mode, in turn, is divided into one-handed mode, two-handed mode, pointer control mode, sequence mode, and calibration mode. As a result, the same basic event can be interpreted in different ways depending on the current interaction scenario.

Since Block 4 is responsible not for image processing but for decision-making logic, it is advisable to present its decomposition in the form of a flowchart. This allows us to illustrate the system activity check, the selection of the current mode, the mode-based

interpretation of the gesture, context-based alignment, and the generation of a high-level command or the reset of temporary states (Fig. 6).

The block diagram details the operational logic of Block A4, “Mode-Context-Based Interpretation of a Gesture Event.” The input for this stage is a basic gesture event, generated based on the coordinates of key points

on the hand and geometric features. In the first step, the system checks whether it is active and whether the received event is valid.

If these conditions are not met, a high-level command is not generated, temporary states are reset, and then the on-screen prompts, event log, and internal system state are updated.

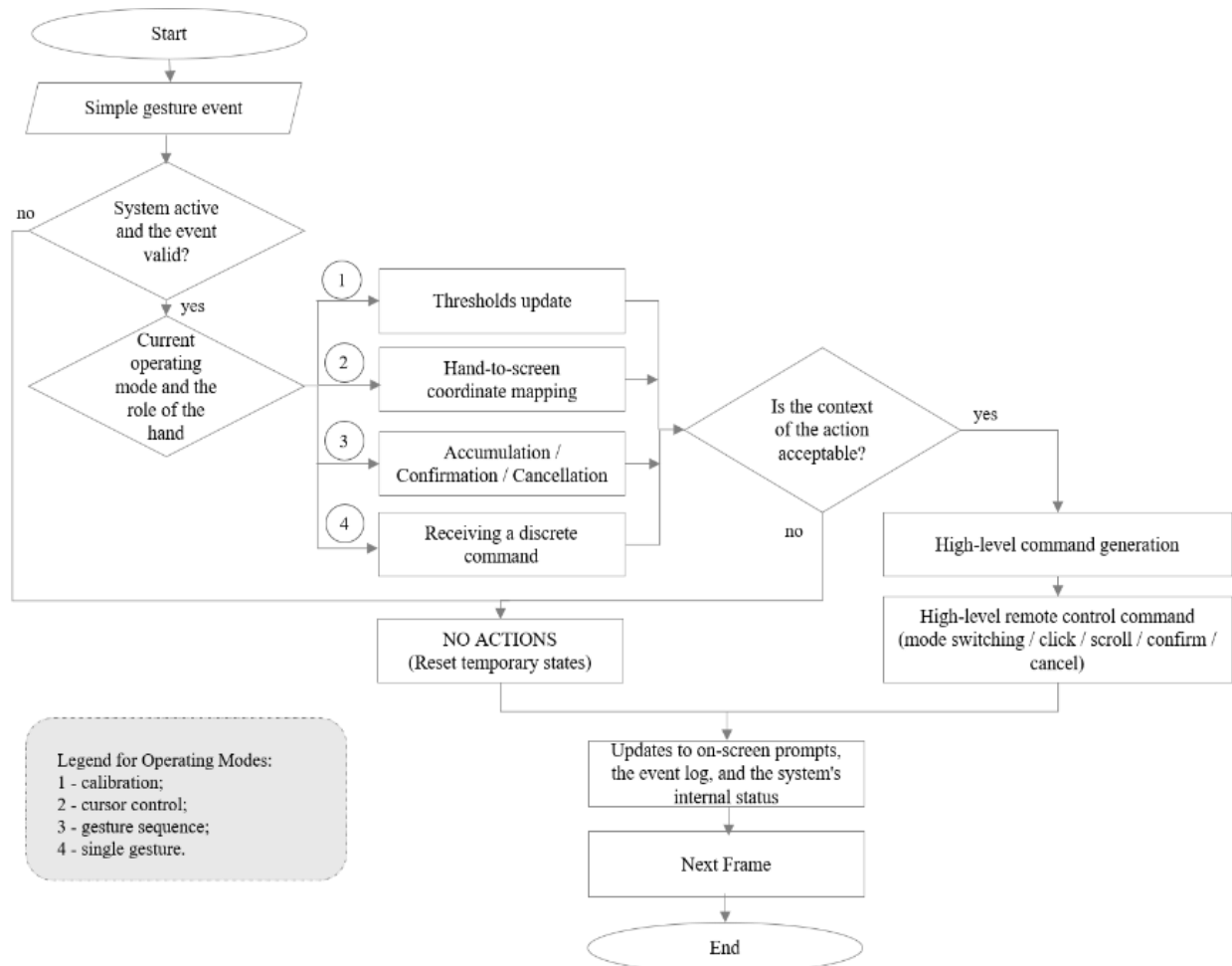


Fig. 6. Flowchart of the Mode-Context Interpretation of a Gesture Event

If a gesture event is valid, the system determines the current operating mode and the role of the hand. The proposed method provides for four main modes: calibration, pointer control, gesture sequence, and single gesture. In calibration mode, threshold values are updated; in pointer control mode, hand coordinates are mapped to screen coordinates; in gesture sequence mode, gestures are used to accumulate, confirm, or cancel a command; and in single-gesture mode, a discrete command is generated.

After the mode interpretation, the system checks whether the action is valid in the current context. At this stage, the system takes into account the previous state of the system, the role of the hand, the active mode, the content type, the interaction scenario, and the presence of confirmation. If the action does not match the current context, the system does not generate a control command. If the action is valid, a high-level remote-control command is generated, such as mode switching, clicking, scrolling, confirmation, or cancellation.

The final step is to update the on-screen prompts, the event log, and the system's internal state, after which the system proceeds to process the next frame. Thus, the flowchart shows that the result of the method is not simply a recognized gesture, but a contextually appropriate remote-control command.

The data presented in Table 3 confirm that the proposed method is operational under the selected experimental conditions and can support the execution of applied interaction scenarios.

The results demonstrate that the system can process visual input, detect gesture events, interpret them according to the current mode and scenario, and generate corresponding control actions.

Therefore, the conducted experiments validate the main idea of the study: contactless interaction with visual and textual content should be based not only on gesture recognition, but also on contextual interpretation of the recognized gesture event.

Table 3 – Summary of the results of the empirical validation of the proposed method

Evaluation Group	Condition / Scenario	Key Indicators
Gesture Recognition Accuracy	Base conditions, 1280×720	Average accuracy – 0,873
	Distance to the camera: 2 m	Average accuracy – 0,673
	Low light, 20 lx	Average accuracy – 0,800
	Change in the hand angle	Average accuracy – 0,513
	Resolution 320×180	Average accuracy – 0,854
	Resolution 1920×1080	Average accuracy – 0,927
Consistency of Individual Gestures	OPEN_PALM	The range of accuracy under different conditions: 0,933 – 0,967, average $\approx$ 0,956
	PINCH	The range of accuracy under different conditions: 0,300 – 0,867, average $\approx$ 0,578
	PINCH_MIDDLE	The range of accuracy under different conditions: 0,333 – 0,867, average $\approx$ 0,650
	FIST	The range of accuracy under different conditions: 0,167 – 1,000, average $\approx$ 0,800
	THUMBS_UP	The range of accuracy under different conditions: 0,733 – 0,967, average $\approx$ 0,878
Scenario Testing	Basic system and navigation actions, using a file operation as an example – creating and opening a folder	Time required for gesture control: 48,692 ms, time spent using the mouse and keyboard – 15,127 ms, errors – 1, critical – 0.
	Working with text or web content using web navigation as an example – opening a browser and navigating to a page	Time required for gesture control – 65129 ms, time spent using the mouse and keyboard – 16206 ms, errors – 0, critical – 0.
	Managing Graphic Objects: An Example Using the Copy and Drag-and-Drop Operations	Time required for gesture control – 80317 ms, time spent using the mouse and keyboard – 13401 ms, errors – 3, critical – 0.
The Effect of Resolution on The Scenarios	320×180, web scenario	Time – 80433 ms, errors – 5, FPS – 22,2
	1280×720, web scenario	Time – 65129 ms, errors – 0, FPS – 19,7
	1920×1080, web scenario	Time – 59379 ms, errors – 2, FPS – 14,4
	1920×1080, drag-and-drop	Time – 58862 ms, errors – 6, critical – 1, FPS – 14,4

This makes it possible to reduce the ambiguity of gesture commands and adapt the system response to the current interaction scenario.

4. Discussion of the Experimental Results of the Contactless Interaction Method

An analysis of the overall results shows that the proposed method provides the most stable performance under standard shooting conditions and when the video stream has sufficient detail. The highest average accuracy was achieved at a resolution of 1920×1080; however, this mode is associated with a higher computational load and a lower frame rate. Therefore, for practical use, the 1280×720 mode is recommended, as it provides an acceptable balance between accuracy, performance, and the number of errors.

Among the individual gestures, OPEN_PALM proved to be the most robust. This is because it has a distinct hand shape and does not require precise determination of small distances between the fingers. The PINCH and PINCH_MIDDLE gestures are more sensitive to image quality, camera angle, and distance from the camera, since their recognition depends on the correct determination of the position of the fingertips. The greatest decrease in accuracy is observed when the

hand's angle changes, which confirms the critical importance of finger overlap for RGB-based tracking.

The scenario testing showed that all the actions considered can be performed using gesture controls. The most stable use case was web browsing, for which no errors were recorded in basic mode. The most challenging scenario turned out to be drag-and-drop, as it requires not only the recognition of a specific gesture but also consistent control while moving an object.

A comparison with traditional input methods shows that gesture control is slower than a mouse and keyboard in terms of execution time. This is to be expected, since the proposed method involves additional steps: hand detection, gesture event recognition, mode verification, contextual interpretation, and command generation. At the same time, the absence of critical errors in basic scenarios confirms the method's viability as a supplementary channel for contactless interaction with visual and text content.

5. Conclusions

This paper proposes a conceptual model of context-adaptive, inclusive user interaction with an information system and the external environment, which takes into account the user's input request, individual needs, the

application scenario, the user's context, the interpretation of input data, the selection of an assistance strategy, and the generation of feedback via a multimodal interface.

This paper presents a method for contactless interaction with visual and text content based on RGB hand tracking and contextual gesture interpretation. The proposed method ensures the sequential conversion of an RGB video stream into an interface control action by identifying key points on the hand, calculating geometric features, forming a basic gesture event, interpreting it in terms of mode and context, and generating a high-level remote-control command.

A functional model of the method has been developed, in which the key element is the block for mode-contextual interpretation of a gesture event. Unlike a direct "gesture-command" mapping, the proposed approach evaluates a gesture event by taking into account the active mode, the system's previous state, the role of the hand, the presence of confirmation, the type of content, and the interaction scenario. This reduces the number of false triggers and increases the predictability of command generation.

The effect of external conditions on gesture recognition accuracy was experimentally evaluated. Under baseline conditions, at a resolution of 1280×720, the average accuracy was 0.873. The highest average accuracy was obtained at a resolution of 1920×1080 – 0.927. H

however, this mode involves a higher computational load; therefore, for practical use, the 1280×720 mode is preferable as a compromise between accuracy, speed, and the number of errors.

It was found that the most critical external factor is a change in the hand's angle, which reduces the average accuracy to 0.513. Increasing the distance to the camera to 2 m also negatively affects recognition quality, reducing the average accuracy to 0.673. Reduced lighting to 20 lux had a more moderate effect: the average accuracy was 0.800. This confirms that for the RGB-based approach, hand visibility, a frontal hand position, and sufficient image detail are particularly important.

A comparison of the stability of individual gestures showed that the OPEN_PALM gesture is the most stable,

with its accuracy varying between 0.933 and 0.967 under different conditions. The PINCH and PINCH_MIDDLE gestures proved to be more sensitive to image quality, distance, and viewing angle, since their recognition depends on the precise determination of the position of the fingertips.

Scenario testing confirmed the effectiveness of the proposed method during basic system operations, web navigation, and control of graphical objects. All basic scenarios were executed without critical errors, confirming the method's suitability as an auxiliary channel for contactless interaction. At the same time, gesture control is slower than a mouse and keyboard in terms of execution time, which is due to the additional steps involved in hand tracking, gesture event recognition, mode verification, contextual interpretation, and command generation.

Further research should focus on improving resistance to changes in hand orientation, enhancing the recognition of subtle and dynamic gestures, adaptively adjusting thresholds for different users, expanding testing scenarios, and deeper integration of gesture control with applications.

Conflict of interest

The authors declare that they have no conflict of interest regarding this research, including financial, personal, authorship-related, or any other type of conflict that could affect the research and its results presented in this article.

Funding

This study was conducted with financial support from the research project "Multimodal Indoor Navigation and Assistive Support System for Users with Special Needs" (state registration number: 0126U002094).

Data availability

The manuscript has no associated data.

Use of artificial intelligence tools

The authors confirm that artificial intelligence technologies were not used in the preparation of this work.

REFERENCES

1. World Health Organization (2023), "Disability", available at: <https://www.who.int/news-room/fact-sheets/detail/disability-and-health>
2. World Health Organization (n.d.), "Hand hygiene", available at: <https://www.who.int/teams/integrated-health-services/infection-prevention-control/hand-hygiene>
3. World Health Organization (2021), "Key facts and figures. World Hand Hygiene Day 2021", available at: <https://www.who.int/campaigns/world-hand-hygiene-day/2021/key-facts-and-figures>
4. Vysocký, A., Grushko, S., Oščádal, P., Kot, T., Babjak, J., Jánoš, R., Sukop, M., and Bobovský, Z. (2020), "Analysis of Precision and Stability of Hand Tracking with Leap Motion Sensor", *Sensors*, vol. 20, no. 15, article 4088, doi: <https://doi.org/10.3390/s20154088>
5. Jeon, H., Choi, H., Noh, D., Kim, T., and Lee, D. (2022), "Wearable Inertial Sensor-Based Hand-Guiding Gestures Recognition Method Robust to Significant Changes in the Body-Alignment of Subject", *Mathematics*, vol. 10, no. 24, article 4753, doi: <https://doi.org/10.3390/math10244753>
6. Yarovy, V., Barkovska, O., Maksymov, D., Ni, Y., and Raptanov, D. (2026), "Methods for recognising hand gestures using computer vision to assist people with movement or speech impairments", *Control, Navigation and Communication Systems*, no. 1(83), pp. 135–141, doi: <https://doi.org/10.26906/SUNZ.2026.1.135>
7. Barkovska, O., Ruban, I., Tymoshenko, D., Holovchenko, O., and Yankovskyi, O. (2025), "Research on mobile machine learning platforms for human gesture recognition in human-machine interaction systems", *Technology Audit and Production Reserves*, vol. 2, no. 2(82), pp. 6–14, doi: <https://doi.org/10.15587/2706-5448.2025.325423>

8. Alsharif, B., Alalwany, E., Ibrahim, A., Mahgoub, I., and Ilyas, M. (2025), "Real-Time American Sign Language Interpretation Using Deep Learning and Keypoint Tracking", *Sensors*, vol. 25, no. 7, article 2138, doi: <https://doi.org/10.3390/s25072138>
9. Yaseen, Kwon, O.-J., Kim, J., Jamil, S., Lee, J., and Ullah, F. (2024), "Next-Gen Dynamic Hand Gesture Recognition: MediaPipe, Inception-v3 and LSTM-Based Enhanced Deep Learning Model", *Electronics*, vol. 13, no. 16, article 3233, doi: <https://doi.org/10.3390/electronics13163233>
10. Chen, I.D.S., Yang, C.-M., Wu, S.-S., Yang, C.-K., Chen, M.-J., Yeh, C.-H., and Lin, Y.-H. (2024), "Continuous Recognition of Teachers' Hand Signals for Students with Attention Deficits", *Algorithms*, vol. 17, no. 7, article 300, doi: <https://doi.org/10.3390/a17070300>
11. Gena, C., Vernerio, F., Mattutino, C., Persichella, G., and Lomanto, L. (2019), "Experimenting Touchless Gestural Interaction for a University Public Web-based Display", *Adjunct Proceedings of the 13th Biannual Conference of the Italian SIGCHI Chapter (CHIItaly 2019)*, Padova, Italy, 5 p., available at: <https://arxiv.org/abs/2011.09749>
12. Drăgoi, M.V., et al. (2025), "An Inclusive Offline Learning Platform Integrating Gesture Recognition and Local AI Models", *Biomimetics*, vol. 10, no. 10, article 693, doi: <https://doi.org/10.3390/biomimetics10100693>
13. Şen, Z., Çil, E.T., and Serif, T. (2024), "Web Camera-Based Gamified Rehabilitation System for Improving Fine Motor Skills: Pilot Study", *International Journal of Digital Waste Engineering*, vol. 1, no. 1, pp. 15–22, doi: <https://doi.org/10.5281/zenodo.14580508>
14. Husna, R., Brata, K.C., Anggraini, I.T., Funabiki, N., Rahmadani, A.A., and Fan, C.-P. (2025), "An Investigation of Hand Gestures for Controlling Video Games in a Rehabilitation Exergame System", *Computers*, vol. 14, no. 1, article 25, doi: <https://doi.org/10.3390/computers14010025>
15. Panagiotou, C., Faliagka, E., Antonopoulos, C.P., and Voros, N. (2025), "Multidisciplinary ML Techniques on Gesture Recognition for People with Disabilities in a Smart Home Environment", *AI*, vol. 6, no. 1, article 17, doi: <https://doi.org/10.3390/ai6010017>
16. Hobbs, T., and Ali, A. (2025), "Smart Home Control Using Real-Time Hand Gesture Recognition and Artificial Intelligence on Raspberry Pi 5", *Electronics*, vol. 14, no. 20, article 3976, doi: <https://doi.org/10.3390/electronics14203976>
17. Lopes, M.C., Silva, P.A., Marengo, L., Vilas Boas, E.C., de Carvalho, J.G.A., Ferreira, C.A., Carvalho, A.L.O., Guimarães, C.V.R., Aquino, G.P., and de Figueiredo, F.A.P. (2026), "GECO: A Real-Time Computer Vision-Assisted Gesture Controller for Advanced IoT Home System", *Sensors*, vol. 26, no. 1, article 61, doi: <https://doi.org/10.3390/s26010061>
18. Wang, X., Shen, H., Yu, H., Guo, J., and Wei, X. (2022), "Hand and Arm Gesture-based Human-Robot Interaction: A Review", *Proceedings of the 6th International Conference on Algorithms, Computing and Systems, Association for Computing Machinery*, New York, article 12, pp. 1–7, doi: <https://doi.org/10.1145/3564982.3564996>
19. Yun, G., Kwak, H., and Kim, D.H. (2024), "Single-Handed Gesture Recognition with RGB Camera for Drone Motion Control", *Applied Sciences*, vol. 14, no. 22, article 10230, doi: <https://doi.org/10.3390/app142210230>
20. Marques, P., Váz, P., Silva, J., Martins, P., and Abbasi, M. (2025), "Real-Time Gesture-Based Hand Landmark Detection for Optimized Mobile Photo Capture and Synchronization", *Electronics*, vol. 14, no. 4, article 704, doi: <https://doi.org/10.3390/electronics14040704>

Received (Надійшла) 14.07.2026

Accepted for publication (Прийнята до друку) 02.09.2026

Publication date (Дата публікації) 25.09.2026

ВІДОМОСТІ ПРО АВТОРІВ / ABOUT THE AUTHORS

Барковська Олесь Юрійвна – кандидат технічних наук, доцент кафедри електронних обчислювальних машин, Харківський національний університет радіоелектроніки, Харків, Україна;

Olesia Barkovska – Candidate of Technical Sciences, Associate Professor at the Department of Electronic Computers, Kharkiv National University of Radio Electronics, Kharkiv, Ukraine;

e-mail: olesia.barkovska@nure.ua; ORCID Author ID: <http://orcid.org/0000-0001-7496-4353>;

Scopus Author ID: <https://www.scopus.com/authid/detail.uri?authorId=24482907700>.

Модель та метод інклюзивної взаємодії з візуальним і текстовим контентом на основі контекстної інтерпретації жестів

О. Ю. Барковська

Анотація. **Актуальність.** Безконтактне жестове керування, як допоміжний канал взаємодії з інтерфейсом, на сьогодні набуло широкого поширення у різних сферах життя, оскільки забезпечує зручний, безпечний і, за необхідності, інклюзивний спосіб взаємодії з візуальним, графічним та текстовим контентом. **Предметом** дослідження є процес безконтактного керування базовими навігаційними сценаріями для роботи з візуальним, текстовим і веб-контентом з використанням методів трекінгу кисті та контекстної інтерпретації жестових подій. **Мета** дослідження полягає в розробленні та експериментальній перевірці методу безконтактної взаємодії з візуальним і текстовим контентом на основі RGB-трекінгу кисті та контекстної інтерпретації жестів для реалізації практичних користувацьких сценаріїв. **Головне завдання** дослідження полягає у розробці концептуальної моделі контекстно-адаптивної інклюзивної взаємодії користувача з інформаційною системою та зовнішнім середовищем. **Методи.** Використано функціональне моделювання, методи комп'ютерного зору, аналіз ключових точок кисті, геометричні правила розпізнавання жестів та експериментальне тестування. **Результати.** Запропоновано метод, у якому жестова подія перетворюється на команду лише після визначення типу та аналізу контексту (технічного та сценарного). Експериментальна перевірка проводилась на п'яти базових жестах і трьох сценаріях. **Висновки.** Отримані результати підтвердили працездатність методу як допоміжного каналу безконтактної взаємодії з візуальним і текстовим контентом.

Ключові слова: безконтактна взаємодія, трекінг, розпізнавання жестів, людино-комп'ютерна взаємодія, MediaPipe Hands, жестове керування, інтерфейс, модель, метод.