Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 57 Issue: 11:2024
DOI: 10.5281/zenodo.14039789

ADVANCED TOUCHLESS HCI: A GESTURE RECOGNITION-BASED MULTILINGUAL VIRTUAL KEYBOARD FOR CHARACTER INPUT

MD. ABDUR RAHIM

Associate Professor and Head, Department of Computer Science and Engineering, Pabna University of Science and Technology, Rajapur, Pabna, Bangladesh. Email: rahim@pust.ac.bd

JUNGPIL SHIN *

Professor, School of Computer Science and Engineering, The University of Aizu, Fukushima, Japan.
*Corresponding Author Email: jpshin@u-aizu.ac.jp;

MD. ABUL HASHEM

M.Sc. in Engineering Degree, Department of Computer Science and Engineering, Pabna University of Science and Technology, Rajapur, Pabna, Bangladesh. Email: hashem.pust@gmail.com

HEMEL SHARKER AKASH

B.Sc. in Engineering Degree, Department of Computer Science and Engineering, Pabna University of Science and Technology, Rajapur, Pabna, Bangladesh. Email: hemelakash.170116@s.pust.ac.bd

MD. IMRAN HOSSAIN

Associate Professor, Department of Information and Communication Engineering, Pabna University of Science and Technology, Rajapur, Pabna, Bangladesh. Email: imran05ice@pust.ac.bd

MD. NAJMUL HOSSAIN

Associate Professor, Department Electrical, Electronic and Communication Engineering, Pabna University of Science and Technology, Rajapur, Pabna, Bangladesh. Email: najmul_eece@pust.ac.bd

ABU SALEH MUSA MIAH

Post-Doc Researcher, School of Computer Science and Engineering, The University of Aizu, Fukushima, Japan. Email: musa@u-aizu.ac.jp

Abstract

Virtual keyboard-based non-touch character input systems present an advanced communication method between humans and computers, offering interaction in challenging environments like industrial settings. Extensive research has explored touch and touchless input methods, including hand gestures, aerial handwriting, sign language recognition, and finger alphabet systems. However, many systems require significant learning and complex processing for accurate character recognition. This reveals the need for more efficient, accessible, and low-overhead solutions in non-touch input technologies. This paper presents a virtual keyboard-based character input system that utilizes hand gesture detection to create a novel touchless human-computer interaction (HCI) interface. The study has two key components: a hand gesture recognition system and a character input method. The system leverages MediaPipe's pre-trained models to accurately detect human body keypoints, enabling mid-air typing through intuitive hand gestures. We calculated the angles and distances between various keypoints to extract the features for gesture recognition. OpenCV is used for data collection, and Pynput facilitates keyboard control. A CronoNet architecture-based model powers the system, translating hand gestures into precise keyboard inputs. The virtual keyboard supports seamless transitions between language layouts, including English and Bengali. It recognizes gestures for commands such as scrolling (up/down), swiping (left/right), thumbs up, and finger tapping for input. The system achieved an average accuracy of 96.54% in gesture recognition and 97.07% in character input, showcasing its superiority over state-of-the-art methods.

Index Terms: Gesture Recognition, Virtual Keyboard Interface, Mediapipe, Crononet Architecture.

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 57 Issue: 11:2024
DOI: 10.5281/zenodo.14039789

1. INTRODUCTION

Human-computer interaction (HCI) techniques are improving to make new technology more usable. A set of assessment criteria should be used to evaluate the performance of an ideal HCI, i.e., accuracy, performance, affordability, sociability, mobility, and usability. However, traditional keyboard-based input systems can cause security issues in performance [1]. As the capabilities of current technology depend on the user's desires, goals, and needs, gesture-based technology is becoming more prevalent. Thus, this paper bridges the gap between humans and computers by developing technical interfaces and designs for evaluating successful HCI, making it more straightforward to utilize technology securely in daily life. HCI approaches are advancing to make current technology more usable. It is a constantly evolving field, representing a new means of communication between people and computers in the modern world [2]. Numerous touch and non-touch devices, such as tablet displays and smartphones, facilitate this interaction. However, touch-based systems pose concerns when the environment is unsafe or harmful, a security risk requiring human intervention, or when the user cannot touch a device [3]. Additionally, users must be in safe locations to operate in industrial settings, food plants, or with robots. Users may be concerned about utilizing touch-based devices for critical applications since hackers could retrieve user data, including biometrics, posing security problems.

Modern advancements in these research fields have propelled our applications to new heights, enabling computers to process 3D graphics effortlessly and ushering us into the era of the Metaverse, characterized by Augmented Reality (AR), Virtual Reality (VR), and Mixed Reality (MR). These technologies create immersive worlds where traditional input devices, such as keyboards and mice, become inadequate because they are designed for two-dimensional user interfaces [4]. The use of head-mounted displays or other immersive equipment necessary to create and interact with these environments often obstructs the use of conventional input devices. In addition to these practical limitations, traditional keyboards and mice pose significant hygiene challenges in specific settings. For instance, maintaining a sterile environment in operating theaters is critical, and shared input devices can become vectors for contamination. Similarly, hygiene is paramount in food production facilities, and using traditional input devices can compromise cleanliness [5]. The COVID-19 pandemic has further highlighted these issues, as shared keyboards and mice can contribute to the spread of pathogens. Beyond hygiene concerns, these devices also present security vulnerabilities [6]. Keyboards can be infected with key-logger malware, which captures and transmits every keystroke to attackers, compromising sensitive information. This is particularly concerning in environments where confidentiality is crucial, such as healthcare, finance, and corporate settings [7]. As we advance into more immersive digital realms, developing and adopting new input methods that address these limitations and vulnerabilities becomes increasingly essential. Innovations such as gesture recognition, voice commands, and brain-computer interfaces are emerging as potential solutions, offering more intuitive and secure ways to interact with digital environments [8]. These technologies enhance user experience and mitigate the hygiene and security issues associated with traditional input devices.

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 57 Issue: 11:2024
DOI: 10.5281/zenodo.14039789

For this reason, hand gesture-based virtual interface technology has been widely used to realize the non-touch system described above. Many experiments have been undertaken in non-touch systems (i.e., hand gesture languages), including VR, HCI, brain-computer interfaces, touch-free writing, aerial handwriting, and sign languages [9-11]. However, there may still be an emphasis on learning conditions and overhead processing for recognition. Consequently, in this paper, we propose a virtual interface-based character input system based on a webcam, which is readily available and widely used so that users do not have to type or write on a keyboard or touchpad. This research aims to create a quick, simple input technique using a hygienic and safe character input system. The main contributions of this paper are as follows:

1. We developed a virtual keyboard interface with a design resembling a traditional on-screen keyboard, allowing users to input characters familiarly and intuitively.
2. Utilizing MediaPipe's pre-trained models, our system accurately detects human body part locations, enabling mid-air typing through intuitive hand gestures. We calculated various measurements, including angle and distance, to enhance gesture recognition. Extensive feature extraction was performed to ensure precise gesture identification for accurate character input.
3. We proposed a robust CronoNet architecture that effectively integrates convolutional and recurrent neural networks to capture spatial and temporal features, optimizing hand gesture detection.
4. The interface allows users to input characters using various gesture functions, including scrolling (up/down), swiping (left/right), thumbs up, and finger tapping. The system demonstrates superior performance when compared to state-of-the-art methods.

The structure of this paper is organized as follows: Section 2 provides a comprehensive review of recent research related to character input systems. Section 3 outlines the methodological framework, detailing the body pose estimation process, data preprocessing, feature extraction, and the proposed CronoNet architecture. Section 4 presents the experimental results and includes an in-depth discussion based on these findings. Finally, Section 5 offers the concluding remarks.

2. RELATED WORKS

Extensive and ongoing research has explored non-touch interfaces utilizing hand gestures, with various methods proposed for gesture recognition across diverse applications. Recent advancements in deep learning and machine learning have notably enhanced classification performance in hand gesture recognition tasks. Here are some recent studies related to character input systems, hand gesture recognition, and the application of machine and deep learning techniques for these purposes. An intelligent noncontact gesture-recognition system integrates a triboelectric touchless sensor (TTS) with a deep-learning-based multilayer perceptron, recognizing 16 gestures with 96.5% accuracy and enabling noncontact robot-controlled throat swab collection [14]. In [15], the

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 57 Issue: 11:2024
DOI: 10.5281/zenodo.14039789

authors proposed a method to enhance accuracy and recognition areas in hand gesture recognition using the CapsNet of a deep neural network and Leap Motion by extracting and preprocessing infrared images and training networks. However, the cost of sensor devices is limited and inaccessible to everyone; however, in this paper, we use readily available webcams found on smartphones, laptops, or webcam devices. In [16], the authors introduced a video-based Finger Writing Virtual Character Recognition System (FVCRS) that allows wireless character input using fingertip movements. It accurately recognizes uppercase (95.3%) and lowercase (98.7%) English alphabets. This system focuses solely on English character recognition and lacks a virtual keyboard, making background noise a significant challenge in accurately recognizing input characters. One study introduced a virtual keyboard that enables text input by tracking fingertip location and hand skin tone with a camera, using a keyboard printed on paper affixed to various surfaces [18]. Another study proposed a similar system, using camera-detected hand gestures to control mouse movements and clicks, with an algorithm mapping mouse and keyboard functions through convex hull flaw detection [19].

Touchless devices have become widespread, particularly after the Covid-19 pandemic. The COVID-19 pandemic has increased the need for contactless biometric authentication systems. This work introduces a novel hand gesture-based sign digit recognition system using a memory-efficient deep learning convolutional neural network. Deployed on a Raspberry Pi 4, it achieves 98.47% accuracy in classifying sign language digits for user authentication [20]. This study explored correlations between gesture usability metrics and qualitative properties, identifying the most efficient gestures for surgeon-computer interaction through statistical analysis and usability testing with neurosurgeons [21]. However, hand gesture recognition and noise removal remain challenging, often leading to performance overhead concerns. Character input systems using flick input usually require users to memorize input methods, which can be difficult [3]. In this paper, we propose a virtual keyboard system that mimics the familiar 'on-screen keyboard' for ease of use.

3. PROPOSED METHODOLOGY

This system detects and recognizes essential body poses, serving as the primary input method to identify specific keys on a virtual keyboard. By integrating data from various body positions, the CronoNet architecture plays a crucial role in interpreting gestures and converting them into character inputs. The system generates the desired characters through precise gesture recognition, enhancing interaction efficiency and intuitiveness. This approach highlights the potential of using body position data and advanced neural networks like CronoNet to create more natural, gesture-based human-computer interactions, offering an innovative alternative to traditional input devices like keyboards and mice. Figure 1 presents the general flow of the virtual character input system.

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 57 Issue: 11:2024
DOI: 10.5281/zenodo.14039789

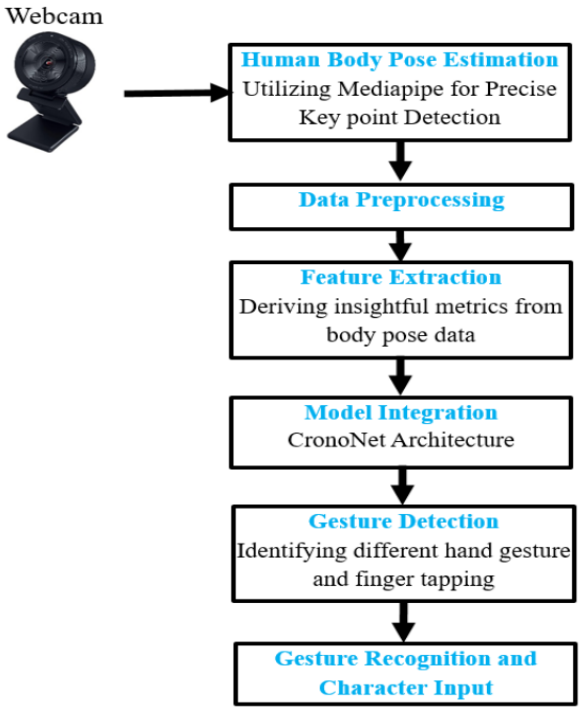

Figure 1: The general process of the virtual character input system

3.1 Dataset Descriptions

This dataset focuses on hand gesture detection and is specifically designed for advanced analysis and training purposes. During data collection, a high-quality camera was used to ensure maximum clarity and fidelity of the captured images. This system concentrated on different gestures from a wide range of possible gestures: scrolling (up/down), swiping (left/right), thumbs up, and finger tapping. To enhance the dataset's versatility, other gestures were combined into a single category, encompassing random or no actions. The dataset comprises 1148 recorded actions, each consisting of a sequence of 30 images, resulting in 34,440 images. This comprehensive repository is meticulously curated to support in-depth analysis, training, and experimentation in hand gesture detection.

Figure 2: Movements of hand gesture from right to left

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 57 Issue: 11:2024
DOI: 10.5281/zenodo.14039789

3.2 Preprocessing

In this research, we used the MediaPipe model for the critical task of human body keypoint detection [12], a key step in preprocessing our dataset. MediaPipe is highly effective in accurately identifying and predicting 32 keypoints across the human body. By utilizing this advanced framework, we carefully selected the keypoints most relevant to our study, ensuring that we extracted the features essential for our analysis. This approach allowed us to focus on the most critical data, enhancing the accuracy and relevance of our findings. Figure 3 shows the body skeleton of different key points.

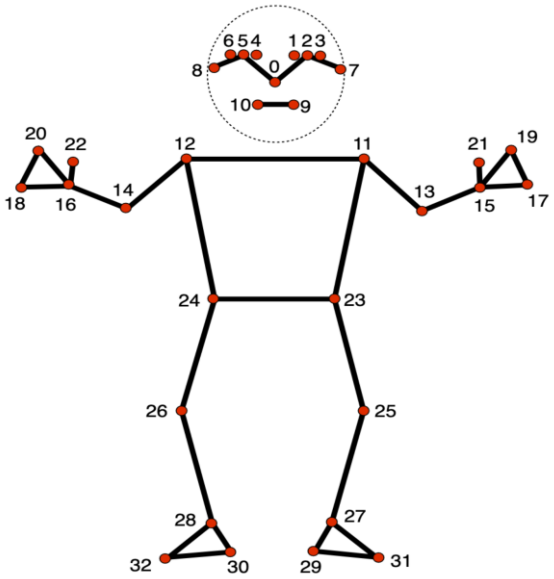

Figure 3: The framework of the body skeleton is made up of different key points

Our preprocessing approach involved a multi-faceted strategy, concentrating on two fundamental techniques:

1. **Distance Measurement**: We meticulously measured the distances between keypoint pairs, enabling a detailed understanding of the spatial relationships within the human body. Table 1 shows the distance point of the different organs of the body skeleton, and Figure 4 shows an example of the distance measurement of other key points of the body skeleton.
2. **Angle Calculation**: We computed the various angles formed between different body parts to advance our investigation. This phase was crucial for understanding the subtleties of postures and body movements. Table 2 represents the different angle representations of the body movements. Furthermore, our approach extended beyond conventional 2D angle computations by incorporating measurements along all three axes: (x, y), (x, z), and (y, z). This enabled a comprehensive evaluation of the body's orientation in three dimensions. By utilizing this thorough preprocessing strategy, we aimed to extract subtle features that capture the essence of human movements, thereby providing a robust foundation for further analysis and model

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 57 Issue: 11:2024
DOI: 10.5281/zenodo.14039789

development in our future research endeavors. Figure 5 shows an example of an angle representation of body movements.

Table 1: Distance between different organs of body keypoints pairs

Distance			
Point-1	**Point-2**	**Organ-1**	**Organ-2**
0	13	Nose	Left Elbow
0	14	Nose	Right Elbow
0	15	Nose	Left Wrist
0	16	Nose	Right Wrist
11	15	Left Shoulder	Left Wrist
12	16	Right Shoulder	Right Wrist
23	13	Left Hip	Left Elbow
24	14	Right Hip	Right Elbow
23	15	Left Hip	Left Wrist
24	16	Right Hip	Right Wrist

Table 2: Angle representation from body movement

Angle					
Point-1	**Point-2**	**Point-3**	**Organ-1**	**Organ-2**	**Organ-3**
0	13	15	Nose	Left Elbow	Left Wrist
0	14	16	Nose	Right Elbow	Right Wrist
11	13	15	Left Shoulder	Left Elbow	Left Wrist
12	14	16	Right Shoulder	Right Elbow	Right Wrist
13	15	21	Left Elbow	Left Wrist	Left Thumb
14	16	22	Right Elbow	Right Wrist	Right Thumb
23	11	13	Left Hip	Left Shoulder	Left Elbow
24	12	14	Right Hip	Right Shoulder	Right Elbow
21	15	19	Left Thumb	Left Wrist	Left Index
22	16	20	Right Thumb	Right Wrist	Right Index
21	15	17	Left Thumb	Left Wrist	Left Pinky
22	16	18	Right Thumb	Right Wrist	Right Pinky
19	15	17	Left Index	Left Wrist	Left Pinky
20	16	18	Right Index	Right Wrist	Right Pinky

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 57 Issue: 11:2024
DOI: 10.5281/zenodo.14039789

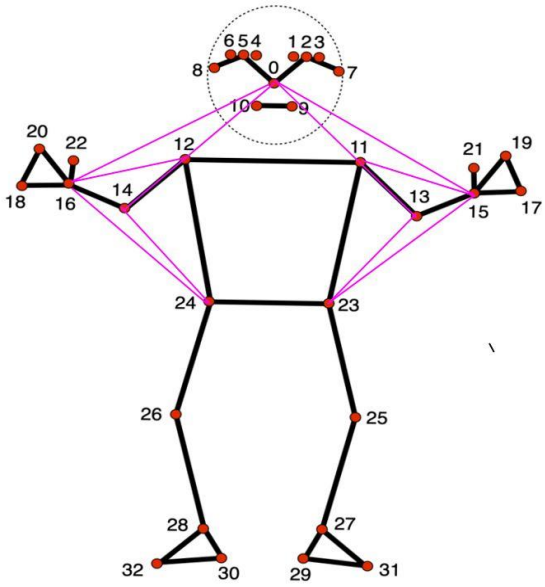

Figure 4: Example of the distance measurement of different keypoints

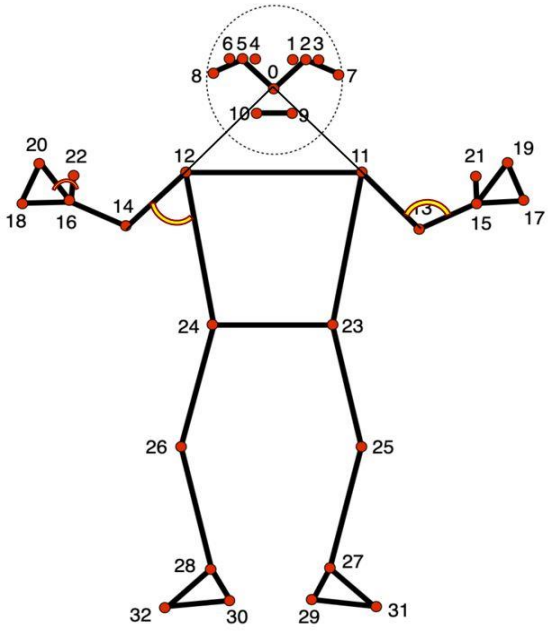

Figure 5: Example of angle representations of body movements

After completing the preprocessing stage using the MediaPipe model, we conducted a careful feature selection process to extract the most significant aspects of each image in our dataset. We concentrated on keypoint positions from a specific subset of keypoints, specifically keypoints 0 and 11-25, which were identified as most relevant for our analysis. By compiling the positional data from these selected keypoints across all images, we created a comprehensive feature set of 114 distinct features for each image. For every

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 57 Issue: 11:2024
DOI: 10.5281/zenodo.14039789

action recorded in our dataset, this feature set is organized in a structured format with dimensions (30, 114), capturing the positional details across the 30 sequential images associated with each action.

3.3 Proposed CronoNet Architecture

The CronoNet architecture, specifically designed for hand gesture detection, integrates convolutional and recurrent neural network components to effectively capture spatial and temporal features from the input data. This section elaborates on the detailed architecture of CronoNet, highlighting its layers and the design rationale. Figure 6 shows the overall flow of CronoNet architecture. The system detects various key points from input images. To extract different features, we calculated angles and distance measurements using the keypoints of the performed gesture. Figure 7 shows the proposed model architecture.

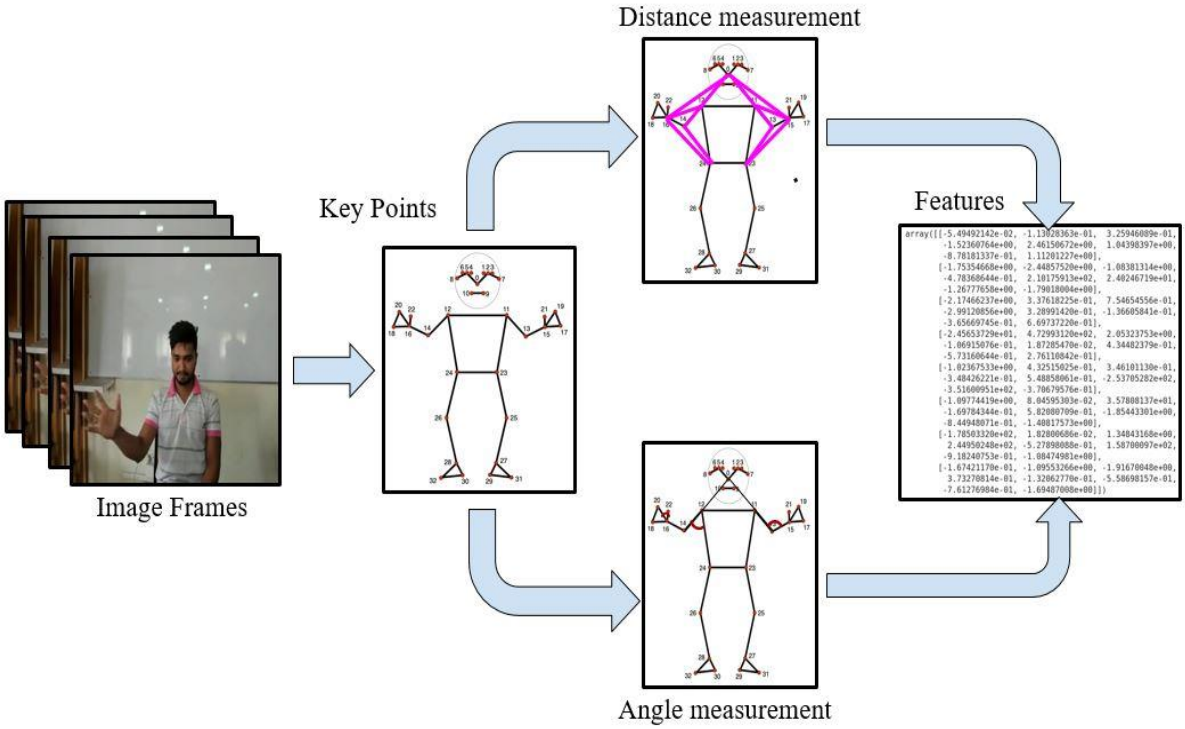

Figure 6: General flow of CronoNet architecture

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 57 Issue: 11:2024
DOI: 10.5281/zenodo.14039789

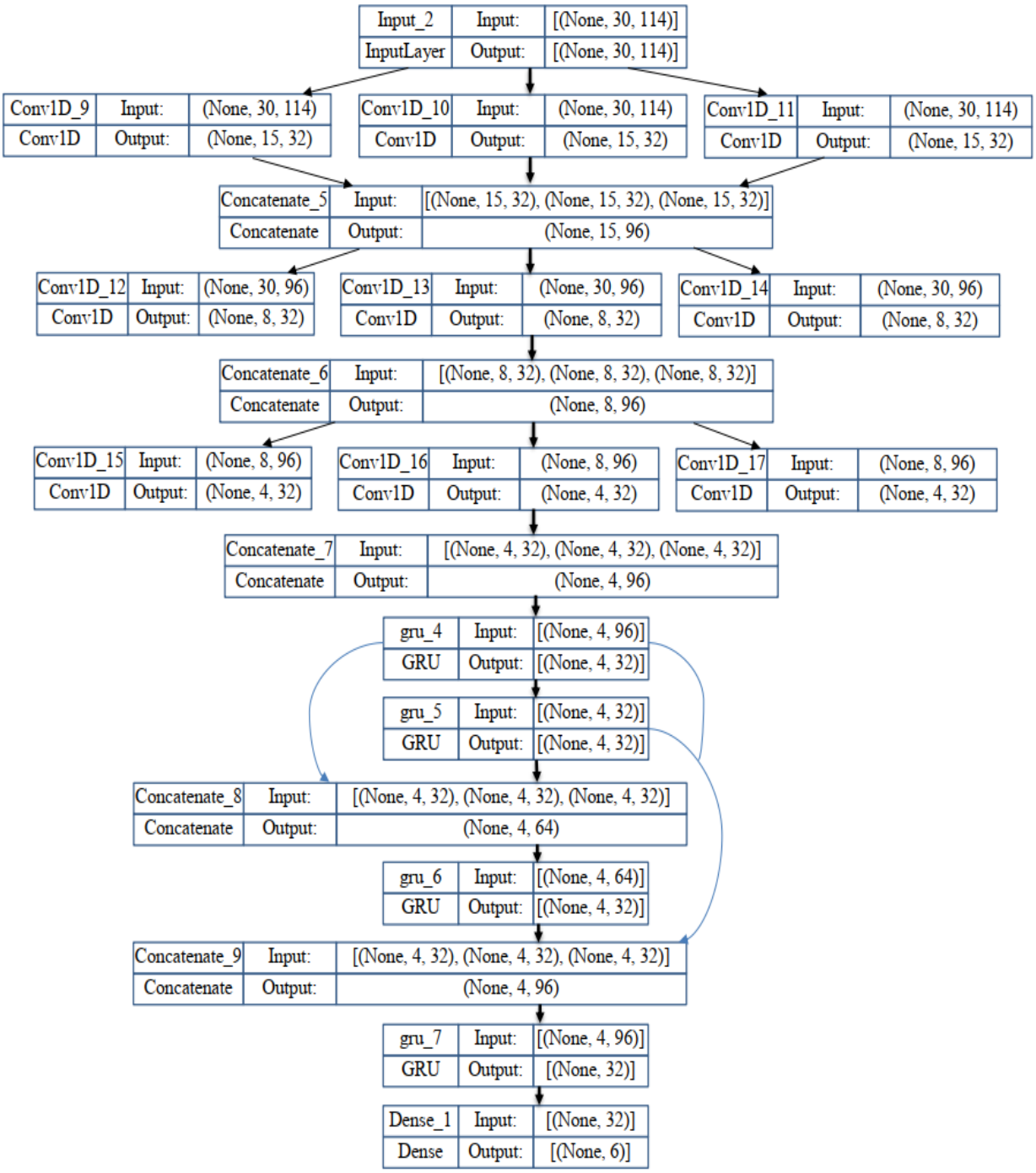

Figure 7: The proposed CronoNet model architecture

The CronoNet architecture comprises a series of convolutional blocks followed by gated recurrent units (GRUs) [13], designed to process sequential skeleton data extracted from video frames.

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 57 Issue: 11:2024
DOI: 10.5281/zenodo.14039789

The key components of the architecture are as follows:

1. Input Layer: The input layer receives sequences of skeleton data points. In our experiment, the input shape was defined as (30, 114), representing 30 time steps and 114 features per time step.
2. Convolutional Blocks: Three convolutional blocks extract spatial features at different scales. Each block consists of three convolutional layers with varying kernel sizes (2, 4, and 8) to capture features at various resolutions. These layers use ReLU activation and are followed by concatenation to merge the features.
   - **First Convolutional Block**: The input is passed through three Conv1D layers with kernel sizes of 2, 4, and 8, each containing 32 filters, and the outputs are concatenated.
   - **Second Convolutional Block**: The concatenated output from the first block is processed through another set of three Conv1D layers, followed by concatenation.
   - **Third Convolutional Block**: This process is repeated, resulting in a final concatenated output representing the spatial features.
3. To capture temporal dependencies, the model integrates a series of GRU (Gated Recurrent Unit) layers:
   - **First GRU Layer**: The output from the convolutional blocks is fed into a GRU layer with 32 units, generating sequences of length 4 with 32 features.
   - **Second GRU Layer**: The output from the first GRU layer is further processed by a second GRU layer, where the output sequences are concatenated with those from the preceding GRU layer.
   - **Third GRU Layer**: An additional GRU layer processes the concatenated output, enhancing the capture of temporal dependencies.
   - **Final GRU Layer**: The final GRU layer reduces the sequence to a single output vector of 32 features.
4. Dense Layer: The final output from the GRU layers is input into a Dense layer with 6 units, corresponding to the number of gesture classes. This Dense layer employs a softmax activation function to generate class probabilities.

3.4 Virtual Keyboard

Figure 8 illustrates the framework of the virtual keyboard system designed for multilingual input, specifically supporting both Bangla and English keyboards. The user can input characters through various hand gestures, with the keyboard layout resembling a conventional computer on-screen keyboard. The virtual keyboard provides two main options: "Show" and "Exit." Selecting the "Show" option activates the corresponding keyboard, allowing the user to input characters via hand gestures. This feature enables users to interact with the system in a manner that mimics typing on a physical keyboard, thereby facilitating a more intuitive and accessible user experience.

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 57 Issue: 11:2024
DOI: 10.5281/zenodo.14039789

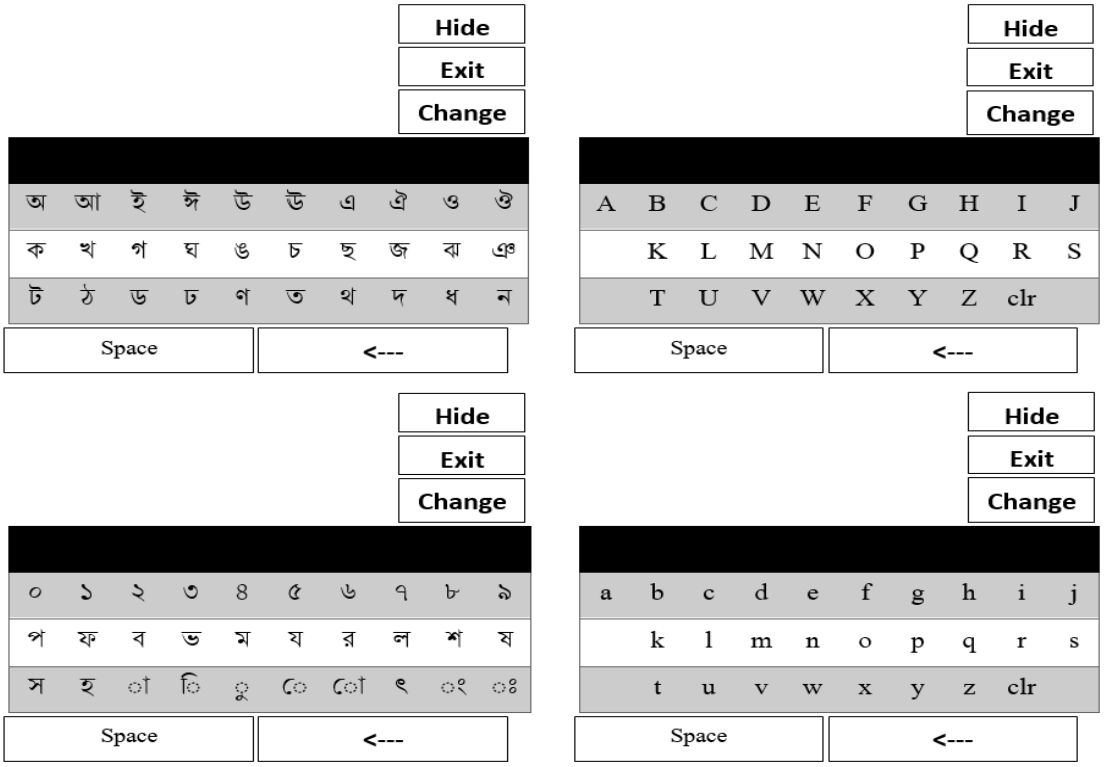

Figure 8: The framework of the virtual keyboard

In addition to the basic input functionality, the virtual keyboard includes a "Change" option, enabling users to switch between the Bangla and English keyboards seamlessly. This feature is particularly beneficial for bilingual users who must alternate between languages. Furthermore, a "Hide" option is available for users who wish to temporarily conceal the keyboard, providing flexibility in managing screen space. The system also integrates a textbox that is displayed alongside the keyboard. This textbox serves as a real-time display of the characters being input, allowing users to monitor their input and make corrections as necessary. Combining these features ensures that the virtual keyboard is both user-friendly and versatile, catering to the needs of a diverse user base.

This research uses various hand gestures for character input in the virtual keyboard system. The "Swipe Left" gesture switches the language input mode by moving the right hand from right to left. A simultaneous up-and-down motion of both hands changes the keyboard layout—either switching to the Bangla keyboard for digit input or toggling between uppercase and lowercase letters on the English keyboard. When the virtual keyboard is not visible, the same up-and-down gestures enable scrolling through the interface. The "Thumbs Up" gesture simulates pressing the "Enter" key. While typing on the virtual keyboard, "Swipe Left" and "Swipe Right" gestures function as "Delete" and "Space," respectively. Finally, a thumb and index finger tapping gesture is used to input a character. Table 3 provides a detailed description of the gestures used for character input in the virtual keyboard.

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 57 Issue: 11:2024
DOI: 10.5281/zenodo.14039789

Table 3: Description of the gestures from character input in the virtual keyboard

Functions	Description
'Swipe Left'	Switches the language mode in Bangla or English keyboard
	Execute the 'Delete' function in the virtual keyboard when the character input is enabled.
'Swipe Right'	Switches the language mode
	Execute the 'Space' function in the virtual keyboard when the character input is enabled.
Thumbs Up	Simulate the "Enter" key.
Scroll Up	Switches the virtual keyboard in digits input
Scroll Down	toggling between uppercase and lowercase letters on the English keyboard
Finger Tapping	Input a character.

4. EXPERIMENTAL ANALYSIS AND DISCUSSION

4.1 Experimental Setup and Data Collection

The experimental data was collected using a video acquisition camera, specifically the Logitech BRIO model, which captures video at 30 frames per second. We used Google Colab for processing, leveraging its specifications of a 2-core processor, 16GB RAM, and a 16GB T4 GPU (x2).

After collecting user video data, we extracted skeletal keypoints from each frame using the MediaPipe model, yielding 32 points across the human body. The computation time for this process ranged between 5-10 milliseconds, which did not impact the video's speed or timing. For the virtual keyboard system, utilizing all the skeletal keypoints was unnecessary.

Instead, we focused on relevant features extracted from the keypoints, specifically those that measured distances to identify hand gestures. Additionally, we calculated angles between key joints such as the shoulder-elbow-wrist and hip-shoulder-elbow. These features were then concatenated, sorted chronologically, and used for gesture recognition in the keyboard system.

4.2 Experimental Results and Analysis

Our research employed the CronoNet framework as the core architecture for hand gesture detection in character input. The gesture recognition model was trained over 100 epochs, with performance metrics carefully monitored throughout. We also trained various models to evaluate our proposed architecture, including XGBoost, Extra Trees, SVC, Multilayer Perceptron (MLP), Passive Aggressive, Ridge CV, Random Forest (RF), Bagging, and K-nearest Neighbors.

Table 4 summarizes the CronoNet model architecture parameters. This architecture, which integrates convolutional and recurrent layers, is specifically designed to efficiently handle the complexities of hand gesture detection by leveraging both spatial and temporal data.

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 57 Issue: 11:2024
DOI: 10.5281/zenodo.14039789

Table 4: Parameter description of the proposed architecture

Layer (type)	Output Shape	Param #	Connected to
input_1 (InputLayer)	(None, 30, 114)	0	[]
conv1d_1 (Conv1D)	(None, 15, 32)	7328	input_2[0][0]
conv1d_2 (Conv1D)	(None, 15, 32)	14624	input_2[0][0]
conv1d_3 (Conv1D)	(None, 15, 32)	29216	input_2[0][0]
concatenate_1 (Concatenate)	(None, 15, 96)	0	conv1d_9[0][0], conv1d_10[0][0], conv1d_11[0][0]
Conv1d_4 (Conv1D)	(None, 8, 32)	6176	concatenate_5[0][0]
conv1d_5 (Conv1D)	(None, 8, 32)	12320	concatenate_5[0][0]
conv1d_6 (Conv1D)	(None, 8, 32)	24608	concatenate_5[0][0]
concatenate_2 (Concatenate)	(None, 8, 96)	0	conv1d_12[0][0], conv1d_13[0][0], conv1d_14[0][0]
conv1d_7 (Conv1D)	(None, 4, 32)	6176	concatenate_6[0][0]
conv1d_8 (Conv1D)	(None, 4, 32)	12320	concatenate_6[0][0]
conv1d_9 (Conv1D)	(None, 4, 32)	24608	concatenate_6[0][0]
concatenate_3 (Concatenate)	(None, 4, 96)	0	conv1d_15[0][0], conv1d_16[0][0], conv1d_17[0][0]
gru_1 (GRU)	(None, 4, 32)	12480	concatenate_7[0][0]
gru_2 (GRU)	(None, 4, 32)	6336	gru_4[0][0]
concatenate_4 (Concatenate)	(None, 4, 64)	0	gru_4[0][0], gru_5[0][0]
gru_3 (GRU)	(None, 4, 32)	9408	concatenate_8[0][0]
concatenate_5 (Concatenate)	(None, 4, 96)	0	gru_4[0][0], gru_5[0][0], gru_6[0][0]
gru_4 (GRU)	(None, 32)	12480	concatenate_9[0][0]
dense_1 (Dense)	(None, 6)	198	gru_7[0][0]
Total params	**178278 (696.40 KB)**	**Trainable params**: 178278 (696.40 KB)	**Non-trainable params**: 0 (0.00 Byte)

To further assess the pperformance of CronoNet, we evaluated its precision, recall, F1 score, and accuracy for each specific task. The detailed results are shown in Table 5.

Table 5: The Precision, Recall, F1-Score, and the accuracy of different hand gestures

Task	Precision	Recall	F1	Accuracy
Thumbs up	0.9688	0.8611	0.9118	0.8611
Scroll Down	0.9756	1	0.9877	1
Scroll Up	0.9512	1	0.975	1
Swipe Left	0.9737	0.9737	0.9737	0.9737
Swipe Right	1	1	1	1
No Action	0.9268	0.95	0.9383	0.95
Average	0.966	0.9641	0.9644	0.9654

CronoNet demonstrated the performance across multiple tasks, achieving high precision, recall, and F1-scores. The average precision, recall, and F1-score were 0.9660, 0.9641, and 0.9644, respectively, with an overall accuracy of 0.9654. Tasks such as "down," " up," and "swipe right" achieved perfect accuracy.

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 57 Issue: 11:2024
DOI: 10.5281/zenodo.14039789

Figure 9 presents the confusion matrix for the CronoNet architecture, visually representing the model's performance across different tasks. The confusion matrix shows that CronoNet effectively distinguishes between tasks with minimal misclassifications. Most tasks have many correctly classified instances, underscoring the model's effectiveness in handling various gestures. Figure 9 presents the loss and accuracy of the proposed architecture. Figure 10(a) illustrates the loss curve, depicting the training and validation loss over 100 epochs. It shows how the model's loss decreased and stabilized, indicating effective learning and convergence. Figure 10(b) presents the accuracy curve, showing the training and validation accuracy over the epochs. This curve highlights the improvement and eventual stabilization of the model's accuracy, confirming its robustness and generalization capability.

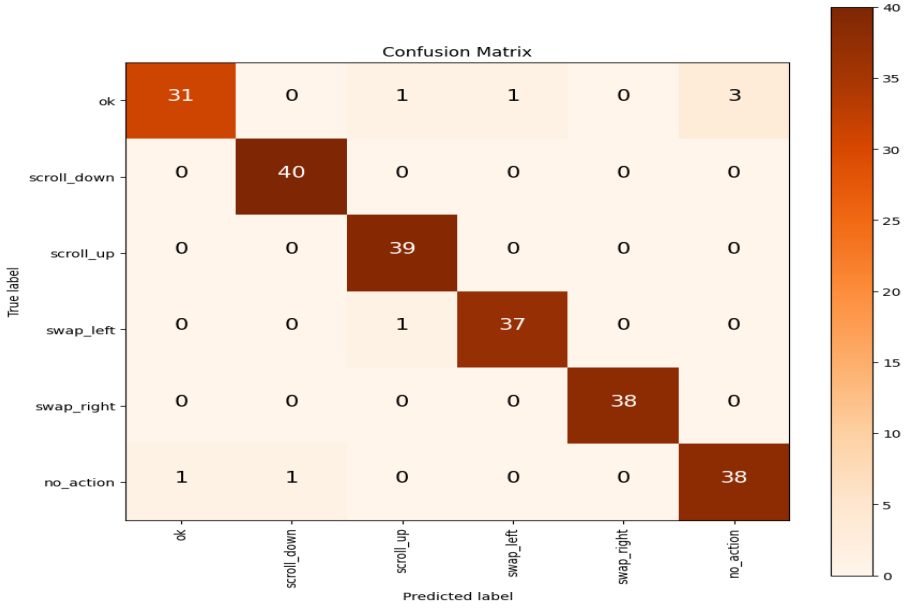

Figure 9: Confusion matrix of the CronoNet Architecture

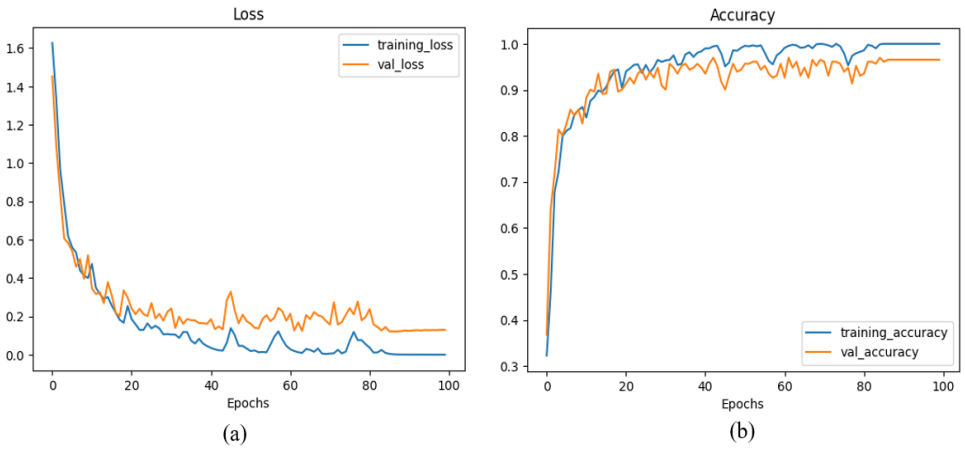

Figure 10: (a) Training and Validation Loss, (b) Training and Validation Accuracy

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 57 Issue: 11:2024
DOI: 10.5281/zenodo.14039789

Furthermore, we evaluated multiple machine learning models on our dataset to assess their performance based on accuracy. Among the models tested, CronoNet, XGBoost, and ExtraTrees demonstrated similar performance, each with an accuracy of 0.9654. This was followed by SVC and Multilayer Perceptron (MLP), which achieved an accuracy of 0.9610. Table 6 shows the performance accuracy of different machine learning models. Moreover, we can say that the CronoNet achieved a better accuracy of 0.9654, outperforming the others.

Table 6: Model Accuracy Comparison

Model	Accuracy
CronoNet	0.9654
XGB	0.9654
Extra Trees	0.9654
SVC	0.961
MLP	0.961
Passive Aggressive	0.9567
Ridge CV	0.9567
Ridge	0.9567
Random Forest	0.9351
Bagging	0.9177
KNeighbors	0.8788

The CronoNet architecture exhibited robust performance across various metrics and tasks, proving a viable choice for gesture recognition. Analysis of the confusion matrix confirmed CronoNet’s ability to classify different tasks with high precision accurately. As illustrated by the loss and accuracy curves, the training process also validated the model’s stability and effectiveness.

Our preference for CronoNet was based on its practical applicability and superior real-world performance observed during extensive testing. While other machine learning models demonstrated commendable accuracy in controlled environments, CronoNet showed unparalleled efficacy in real-world scenarios.

The framework's robustness and adaptability were impressive, as it handled environmental factors and noise inherent in real-world data with remarkable resilience. CronoNet's ability to generalize beyond the training dataset and capture temporal dependencies in sequential data suited it exceptionally for hand gesture detection.

Moreover, CronoNet outperformed in scenarios where temporal information was critical, translating into enhanced accuracy and reliability. This superiority underscores its practical utility and effectiveness over traditional machine-learning approaches.

4.3 Character input system

The user inputs characters through various hand gesture functions. Initially, the user activates the virtual keyboard and selects the desired language for input. Users can switch between Bangla and English modes by moving their right hand from right to left or left to right.

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 57 Issue: 11:2024
DOI: 10.5281/zenodo.14039789

To input a character, the user navigates over the keyboard, selects the target character, and taps with their index finger and thumb. The character is then displayed in a textbox on the virtual keyboard. While the system achieves 100% accuracy in character recognition, its overall precision decreases when users accidentally select incorrect characters. Figure 11 illustrates the character input system.

Random participants entered 41 characters, 34 in Bangla and 17 in English. In a Bangla sentence, the user inputs something like "আমাদের দেশের নাম বাংলাদেশ," and for English, it would be "I love my country." The input of Bangla and English characters is shown in Figure 12. The average accuracy of character selection and the recognition rate for all users are shown in Figure 13.

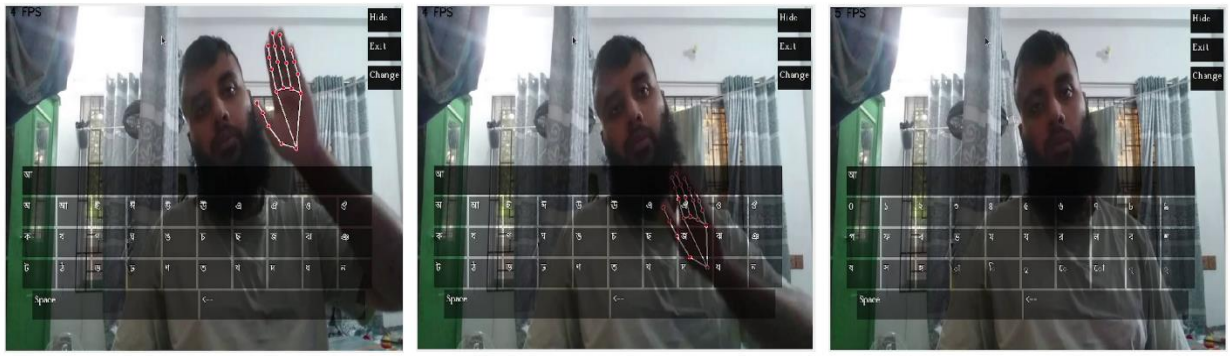

Figure 11: Example of the virtual keyboard for character input

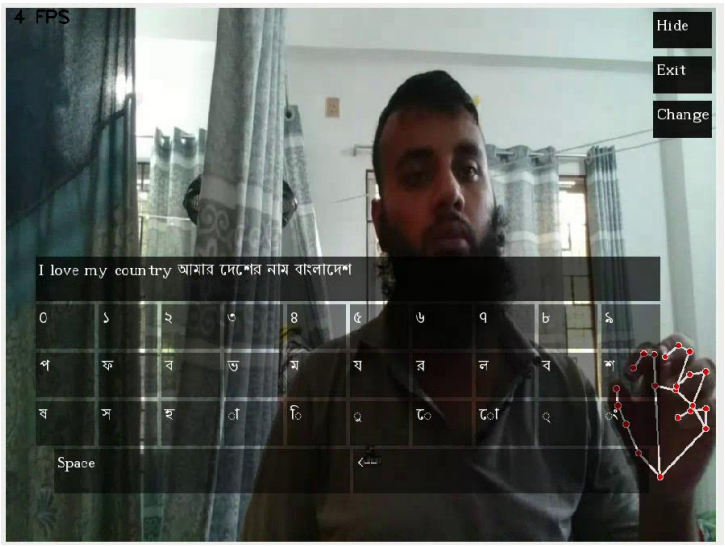

Figure 12: The example of the random respondent of character input

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 57 Issue: 11:2024
DOI: 10.5281/zenodo.14039789

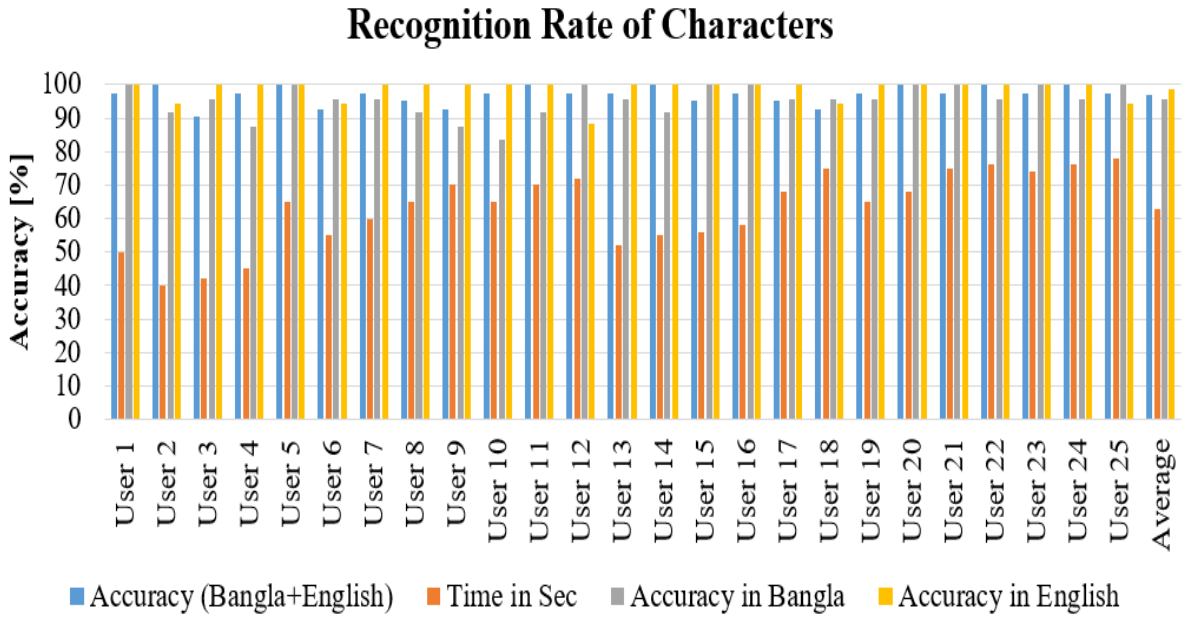

Figure 13: The average recognition rate of the character input system

A total of 25 random respondents were asked to perform the character input task. Figure 12 shows that the average recognition accuracy for inputting Bangla and English characters is 97.07%. Specifically, the accuracy for Bangla characters is 95.5%, while for English characters, it is 98.59%. The average time to input the total number of characters is 63 seconds. Consequently, users can input an average of 39.04 characters per minute.

5. CONCLUSION

This paper presented a virtual keyboard system that uses hand gesture recognition to improve Human-Computer Interaction (HCI). Using MediaPipe's pre-trained models, our system accurately detects human body parts, allowing users to type in mid-air. This touchless interface enhances accessibility for those with physical impairments and provides a hygienic alternative to traditional keyboards. Using OpenCV for data acquisition and Pynput for keyboard control, we developed a highly efficient system based on the CronoNet architecture. Our model, trained over 100 epochs, achieved a 96.54% accuracy rate and can switch between multiple language layouts, including English and Bangla.

The average character input accuracy is 97.07%, with English character input accuracy at 98.59%. Despite the higher accuracy of the Logistic Regression model in controlled conditions, we chose CronoNet for its robustness and adaptability in real-world scenarios. The CronoNet model's ability to handle environmental noise and capture temporal dependencies made it ideal for practical applications like hand gesture detection. Our research suggests that this Bangla virtual keyboard, with its high accuracy and multilingual support, is a significant advancement in HCI, offering a streamlined and

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 57 Issue: 11:2024
DOI: 10.5281/zenodo.14039789

efficient method for digital communication. Future work will improve accuracy, expand to other languages, and integrate with augmented and virtual reality systems. This research highlights the potential of gesture-based interfaces and paves the way for future innovations in touchless computing.

Acknowledgment

The Research Support and Publication Division of the Bangladesh University Grants Commission (UGC) FY 2022-2023 as part of a research project on information and technology, supported this research.

References

1) M. A. Rahim, J. Shin, and M. R. Islam, "Hand gesture recognition-based non-touch character writing system on a virtual keyboard," Multimedia Tools and Applications, vol. 79, no. 17, pp. 11813-11836, January 2020. https://doi.org/10.1007/s11042-019-08448-6

2) M. Nazar, M. M. Alam, E. Yafi, and M. M. Su'ud, "A systematic review of human-computer interaction and explainable artificial intelligence in healthcare with artificial intelligence techniques," IEEE Access, vol. 9, pp. 153316-153348, November 2021. https://doi.org/10.1109/ACCESS.2021.3127881

3) J. Shin, J. and C. M. Kim, "Non-touch character input system based on hand tapping gestures using Kinect sensor," IEEE Access, vol. 5, pp.10496-10505, May 2017. https://doi.org/10.1109/ACCESS.2017.2703783

4) J. Pognon, J. Chi, A. Salabert, K. Kim, and S.J. Kim, "Meta-Analysis of Global Activities in Augmented Reality (AR) and Virtual Reality (VR)," Augmented Reality and Virtual Reality: Changing Realities in a Dynamic World, pp.335-347, March 2020. https://doi.org/10.1007/978-3-030-37869-1_27

5) K.S. Kyaw, S. C. Adegoke, C.K. Ajani, O. F. Nwabor, and H. Onyeaka, "Toward in-process technology-aided automation for enhanced microbial food safety and quality assurance in milk and beverages processing," Critical reviews in food science and nutrition, vol. 64, no. 6, pp.1715-1735, February 2024. https://doi.org/10.1080/10408398.2022.2118660

6) V. Donne, and M. A. Hansen, "This Isn't Science Fiction: Technology Use During and Post-COVID for Students with Disabilities," Journal of Educational Technology Systems, vol. 0, no. 0, August 2024. https://doi.org/10.1177/00472395241267713

7) M. A. Rahim, J. Shin, M. R. Islam, "Hand gesture recognition-based non-touch character writing system on a virtual keyboard," Multimedia Tools and Applications (Springer), vol. 79, pp. 11813-11836, January 2020. https://doi.org/10.1007/s11042-019-08448-6

8) S. Verma, A. Aqle, "A Review on Roles of Next Generation User Interface to Support People with Disabilities," Nafath. Vol. 9, no. 26, August 2024. https://doi.org/10.54455/MCN2604

9) J. Qi, L. Ma, Z. Cui, Y. Yu, "Computer vision-based hand gesture recognition for human-robot interaction: a review," Complex & Intelligent Systems, vol. 10, no. 1, pp. 1581-606, February 2024. https://doi.org/10.1007/s40747-023-01173-6

10) C. Guger, N. F. Ince, M. Korostenskaja, B.Z. Allison, "Brain-Computer Interface Research: A State-of-the-Art Summary 11," In Brain-Computer Interface Research: A State-of-the-Art Summary, pp. 1-11, January 2024, Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-49457-4_1

11) S. Kapoor, A. Sharma, A. Verma, and S. Singh, "Aeriform in-action: A novel dataset for human action recognition in aerial videos," Pattern Recognition, vol. 140, p.109505, September 2023 https://doi.org/10.1016/j.patcog.2023.109505

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 57 Issue: 11:2024
DOI: 10.5281/zenodo.14039789

12) Sánchez-Brizuela, G., Cisnal, A., de la Fuente-López, E., Fraile, J.C. and Pérez-Turiel, J., 2023. Lightweight real-time hand segmentation leveraging MediaPipe landmark detection. Virtual Reality, 27(4), pp.3125-3132. https://doi.org/10.1007/s10055-023-00858-0

13) S. Farah, N. Humaira, Z. Aneela, and E. Steffen, "Short-term multi-hour ahead country-wide wind power prediction for Germany using gated recurrent unit deep learning," Renewable and Sustainable Energy Reviews, vol. 167, p.112700, October 2022. https://doi.org/10.1016/j.rser.2022.112700

14) H. Zhou, W. Huang, Z. Xiao, S. Zhang, W. Li, J. Hu, T. Feng, J. Wu, P. Zhu, P. and Y. Mao, "Deep-learning-assisted noncontact gesture-recognition system for touchless human-machine interfaces," Advanced Functional Materials, vol. 32,no. 49, pp.2208271, December 2022. https://doi.org/10.1002/adfm.202208271

15) A.R. Lee, Y. Cho, S. Jin, and N. Kim, "Enhancement of surgical hand gesture recognition using a capsule network for a contactless interface in the operating room," Computer methods and programs in biomedicine, vol. 190, p.105385, July 2020. https://doi.org/10.1016/j.cmpb.2020.105385

16) L. Jin , D. Yang, L. X. Zhen, J. C. Huang, "A novel vision-based finger-writing character recognition system," Journal of Circuits, Systems, and Computers. vol. 16, no. 03, pp. 421-36, Jun 2007. https://doi.org/10.1142/S0218126607003757

17) M. Oudah, A. Al-Naji, J. Chahl, "Hand gesture recognition based on computer vision: a review of techniques," Journal of Imaging, vol. 6, no. 8, pp. 73, July 2023. https://doi.org/10.3390/jimaging6080073

18) Y, Zhang, W. Yan, A. Narayanan, "A virtual keyboard implementation based on finger recognition," In 2017 International Conference on Image and Vision Computing New Zealand (IVCNZ), pp. 1-6, December 2017, IEEE. https://doi.org/10.1109/IVCNZ.2017.8402452

19) B.R. Sandhya, C. Amrutha, and S. Ashika, "Gesture Recognition Based Virtual Mouse and Keyboard," In International Conference on Advances in Communication Technology and Computer Engineering, pp. 25-36, February 2023, Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-37164-6_3

20) M. YacinSikkandar, "Design a contactless authentication system using hand gestures technique in COVID-19 panic situation," Annals of the Romanian Society for Cell Biology, vol. 25, no. 1, pp. 2149-2159, 2021.

21) N. Madapana, D. Chanci, G. Gonzalez, L. Zhang, and J. P. Wachs, "Touchless interfaces in the operating room: a study in gesture preferences," International Journal of Human-Computer Interaction, vol. 39, no. 3, pp. 438-448, February 2023. https://doi.org/10.1080/10447318.2022.2041896