sensors

MDPI

Article

Intelligent Gesture Recognition Based on Screen Reflectance Multi-Band Spectral Features

Peiying Lin [1,*], Chenrui Li [2], Sijie Chen [2], Jiangtao Huangfu [2] and Wei Yuan [1]

[1] School of Electrical and Information Engineering, Jiangsu University of Science and Technology, Zhangjiagang 215600, China; yuanwei@just.edu.cn
[2] Laboratory of Applied Research on Electromagnetics, Zhejiang University, Hangzhou 310027, China; 3210103050@zju.edu.cn (C.L.); 22331025@zju.edu.cn (S.C.); huangfujt@zju.edu.cn (J.H.)
* Correspondence: linpeiying@just.edu.cn

Abstract: Human–computer interaction (HCI) with screens through gestures is a pivotal method amidst the digitalization trend. In this work, a gesture recognition method is proposed that combines multi-band spectral features with spatial characteristics of screen-reflected light. Based on the method, a red-green-blue (RGB) three-channel spectral gesture recognition system has been developed, composed of a display screen integrated with narrowband spectral receivers as the hardware setup. During system operation, emitted light from the screen is reflected by gestures and received by the narrowband spectral receivers. These receivers at various locations are tasked with capturing multiple narrowband spectra and converting them into light-intensity series. The availability of multi-narrowband spectral data integrates multidimensional features from frequency and spatial domains, enhancing classification capabilities. Based on the RGB three-channel spectral features, this work formulates an RGB multi-channel convolutional neural network long short-term memory (CNN-LSTM) gesture recognition model. It achieves accuracies of 99.93% in darkness and 99.89% in illuminated conditions. This indicates the system's capability for stable operation across different lighting conditions and accurate interaction. The intelligent gesture recognition method can be widely applied for interactive purposes on various screens such as computers and mobile phones, facilitating more convenient and precise HCI.

Keywords: multi-band spectra; human–computer interaction; gesture recognition

Citation: Lin, P.; Li, C.; Chen, S.; Huangfu, J.; Yuan, W. Intelligent Gesture Recognition Based on Screen Reflectance Multi-Band Spectral Features. *Sensors* **2024**, *24*, 5519. https://doi.org/10.3390/s24175519

Academic Editor: Eui Chul Lee

Received: 12 July 2024
Revised: 22 August 2024
Accepted: 24 August 2024
Published: 26 August 2024

1. Introduction

In contemporary society, digitization has emerged as a crucial trend, fundamentally transforming social dynamics. The widespread development of digital technologies and the growing presence of smart devices are seamlessly integrating into our daily lives [1]. This transformation is facilitated by HCI, which is a discipline focused on designing, evaluating, and implementing interactive computing systems for human use as well as studying the fundamental phenomena [2]. Due to its close collaboration and interaction with users, HCI has become a core area for enhancing the usability of digital devices [3]. Central to visualizing information in this interaction is the display device, which plays a crucial role in data communication and allows for intuitive user interactions [4].

Traditional methods of interaction with display devices rely on desktop setups equipped with keyboards and mice [5]. With technological advancements, more flexible and convenient methods have been adopted. Touch interaction is widely embraced due to its direct method of intuitive control and data transmission [6], enabling the transfer of complex information such as multi-touch [7,8] or multi-user collaboration [9]. Compared with tactile modes, voice interaction eliminates the need for direct physical contact. For instance, in driving scenarios, non-contact voice interaction proves more user-friendly [10], thereby enhancing user satisfaction [11]. However, voice input prolongs interaction response

time [12] and is constrained in environments with background noise [13]. Computer vision enables diverse interactions involving facial [14] and bodily gestures [15], but its accuracy depends on the resolution and frame rate of the camera [16]. According to the World Health Organization (2024), over 466 million people worldwide suffer from severe hearing loss. Gesture-based interaction offers a promising solution for enhancing communication for these individuals [17]. Technologies such as computer vision [18–20], audio [21], and radar detection [22,23] enable gesture-based screen interactions. Feature extraction combined with detection [18–34] is commonly used in gesture recognition. These features include frequency [21], motion [23], skin color [26], skeletal structure [27], and shape [28], as well as spatio-temporal features [29–31] derived from deep networks. Additionally, depth information [32,33] and optical flow [34] are frequently utilized to supplement image data, although this demands more advanced equipment. By combining various features and employing multi-stream techniques, it is possible to achieve more effective feature fusion [35].

The method of light-signal-based interaction offers an alternative non-contact solution. For instance, in medical applications [36], touch-based interaction screens increase the risk of surgical infections. Furthermore, visual and voice interactions require the collection of biological information, which compromises privacy and security. Therefore, infrared laser positioning can be employed as an alternative. Infrared spectra can also be specifically applied in human signal measurement [37]. In addition to infrared technology, industry and the research community have developed numerous visible-light positioning (VLP) systems [38] and visible-light sensing (VLS) systems [39], which require commonly used light emitting diodes (LED) as lighting sources and light sensors to form the systems [40]. Similarly, utilizing visible light for screen sensing involves using ambient light sensors to capture light intensity information from external light sources at various angles relative to the screen [41]. Combining light sensing with gesture interaction provides a convenient and secure method for non-contact interaction [42,43].

This study introduces a gesture recognition approach that combines multi-band spectral features with the spatial characteristics of screen-reflected light. In this approach, display screens are used as light sources for illumination, and various gestures produce unique patterns of reflected spectra in front of the screen. Concurrently, multiple narrowband spectral receivers capture data across multi-band spectra. This combination of spectral data is fused with spatial information, enabling the formation of comprehensive multidimensional features essential for accurate gesture recognition. One of the key advantages of this screen interactive system is its independence from the additional light sources, radar systems, or camera devices commonly used for similar purposes. Moreover, the implementation of cost-effective narrowband spectral receivers enhances affordability without compromising performance. Additionally, this approach addresses privacy concerns by minimizing the collection of biometric information, ensuring a secure and user-friendly interaction environment.

The remainder of the paper is structured as follows: Section 2 introduces a gesture recognition method based on multi-band spectral features, implementing an RGB three-channel narrowband spectral gesture recognition system. Section 3 outlines the data collection process. Section 4 details the RGB multi-channel CNN-LSTM gesture recognition model. Then, the experimental results are presented and discussed in Section 5. Finally, Section 6 serves as the conclusion of this paper.

2. Principles and System

2.1. Principles

A gesture recognition method based on multi-band spectral features is proposed in this work, which combines the spectral and spatial characteristics of screen light reflected from gestures. The specific process is illustrated in Figure 1. The intelligent gesture recognition system according to this method mainly consists of a light-emitting display screen and a plurality of narrowband spectral receivers. The system works by orienting the target gesture toward the screen, in which the screen serves the purpose of providing illumination

on the gesture while displaying normally. The light information reflected by the gesture is captured by multiple narrowband spectral receivers mounted on the screen. These receivers are installed at different positions on the plane where the screen is located, and the reflected light from the gestures generates different spectral distributions in various spatial locations. Furthermore, these receivers capture narrowband spectra from different bands and convert them into photonic signals to obtain light-intensity measurements. As a result, spectral data containing various bands from different coordinates can be received, which provides the possibility to train different characteristics in frequency and spatial domains. Based on the measurements from multiple receivers and combined with classification algorithms, different gestures can be effectively classified, significantly improving classification efficiency and recognition accuracy.

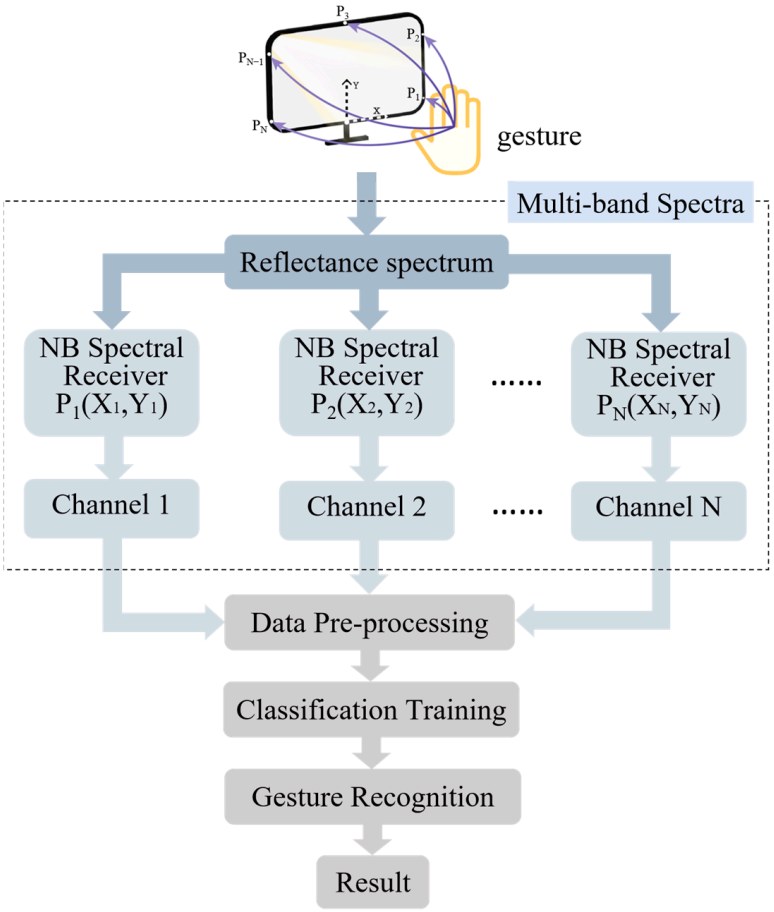

Figure 1. Flowchart of the gesture recognition method based on multi-band spectral features.

2.2. System

According to the method, an RGB three-channel narrowband spectral gesture recognition system is realized as shown in Figure 2a, with three narrowband spectral receivers installed at different coordinates of the screen plane. The light emitted from the screen is reflected by the gestures and then captured by receivers positioned at three coordinates on the screen: bottom-right, bottom-left, and top-center. These receivers record narrowband spectral data corresponding to the red, green, and blue channels, as in Figure 2b. Consequently, the three-channel data incorporate both spectral and spatial information features.

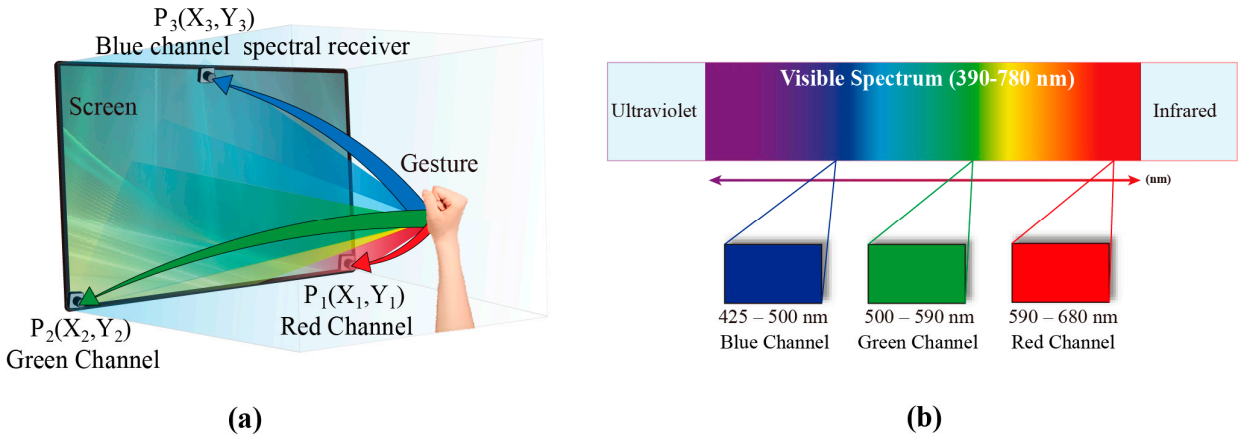

Figure 2. (**a**) RGB three-channel narrowband spectral gesture recognition system; (**b**) filtered RGB three-channel narrowband spectra from the system.

The system configuration is set up as in Figure 3a, with a screen size of 17.2 inches. The narrowband spectral receiver consists of a light sensor and a filter film. The light sensor chip is OPT4001, with a measurement range of 1–918 lux and an accuracy of up to 112 millilux. The data sampling rate is set at 100 Hz in the experiments. The filter film is placed in front of the light sensor to selectively receive specific wavelength light. Figure 4 illustrates the spectral filtering effect of the filter films measured by the spectrometer on the screen light, with Figure 4a depicting the measured spectrum when the screen emits white light. Figure 4b–d depict the corresponding RGB narrowband spectra after passing through the filter films. The RGB spectral wavelengths received through the filter films are 590–680 nm, 500–590 nm, and 425–500 nm, respectively. The filter effectively filters out spectra outside the narrow band without affecting the shape and characteristics of the target narrowband spectra, while also reducing the intensity of the entire spectrum. The use of light-intensity sensors as spectral receivers not only enhances the sensitivity of light detection but also provides convenience and cost reduction compared with spectrometers. The spectral receiver outputs a time series of integrated light intensity corresponding to each narrowband spectrum.

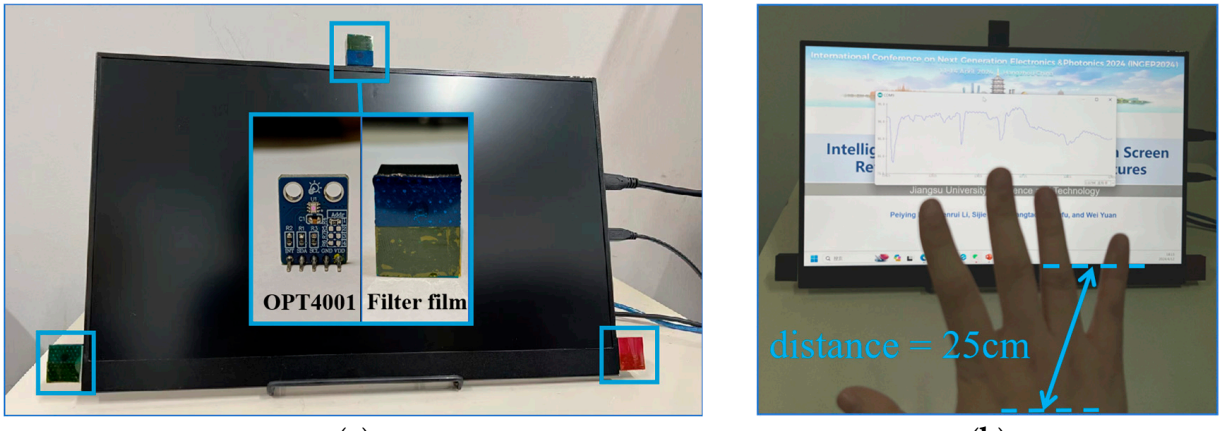

Figure 3. (**a**) The photograph of the system configuration; (**b**) the data fluctuations of the light intensity reflecting hand gesture variation.

During system operation, the display screen emits light normally. When a person's hand is placed within a distance range of 10–70 cm directly in front of the screen, the screen light reflected by gestures reaches the narrowband spectral receivers positioned at different spatial locations. After passing through the filter film, only light of specific

wavelengths is allowed to be received by the light-intensity sensor. The receiver converts narrowband spectral information into light-intensity time series, which are transmitted to the screen control terminal. The intensity information is visualized in real time on the screen. Variations in gestures cause changes in the reflected light intensity, which can be observed as corresponding data fluctuations on the screen in Figure 3b, reflecting changes in hand movements.

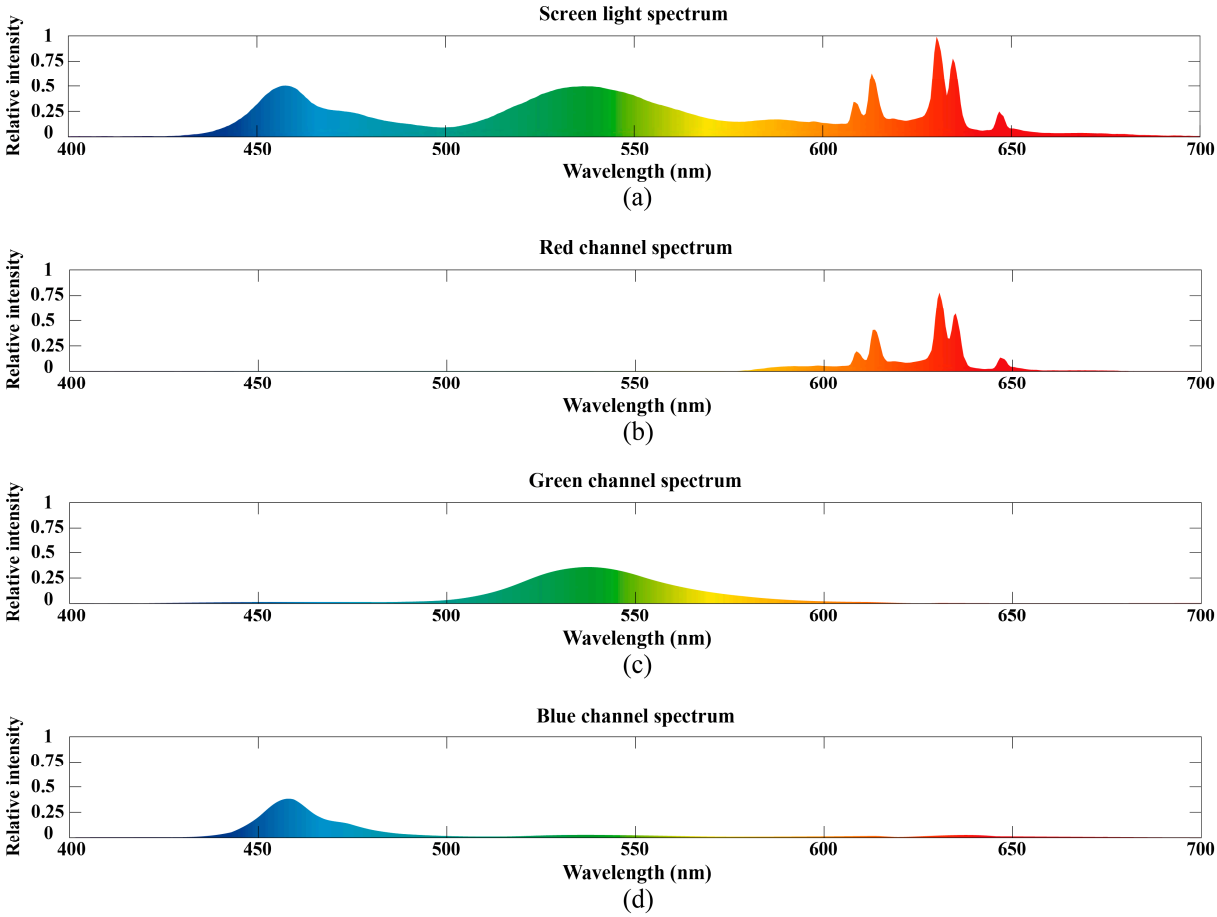

Figure 4. Spectral filtering effects of the filter films on display light measured by the spectrometer: (**a**) The spectrum measured by the spectrometer when the screen emits white light; (**b**) the spectrum through the red channel narrowband filter; (**c**) the spectrum through the green channel narrowband filter; (**d**) the spectrum through the blue channel narrowband filter.

3. Data Collection

Section 2 of the system is designed for implementing gesture-based HCI, applied in eight gestures as depicted in Figure 5, each annotated with a distinct color. The term "Background" refers to the scenario where no gesture is present in front of the screen, serving as a baseline control. The process of data collection is conducted through the RGB three-channel spectral receivers of the system setup, with the datasets structured into two main groups, labeled Dataset 1 and Dataset 2.

In Dataset 1, data collection involved performing eight gestures directly facing the screen in darkness, with the screen display being the only light source. Under identical display conditions, Figure 6 illustrates the light intensity data from RGB three-channel spectral receivers for the eight gestures and control group. In this dataset, variations in light intensity stem from changes in the distribution of screen-reflected light caused by the gestures. The lowest light intensity occurs when no gesture is present, which aligns with the operational principle of the system. Moreover, Figure 6 presents notable differences in data distribution across different channels receiving the same display content, indicating

varying impacts of gestures on the light-intensity information received by each channel. These differences arise from spatial information and spectral wavelength variations across the channels. The multi-channel data constitute multidimensional time-series features, essential for accurate gesture recognition.

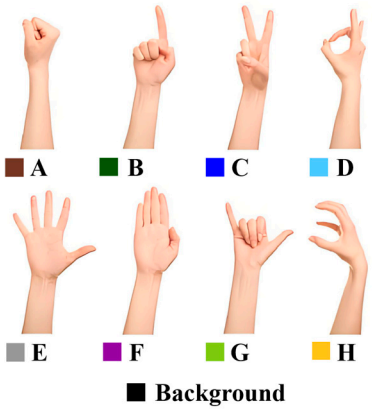

Figure 5. Eight gestures in the gesture recognition.

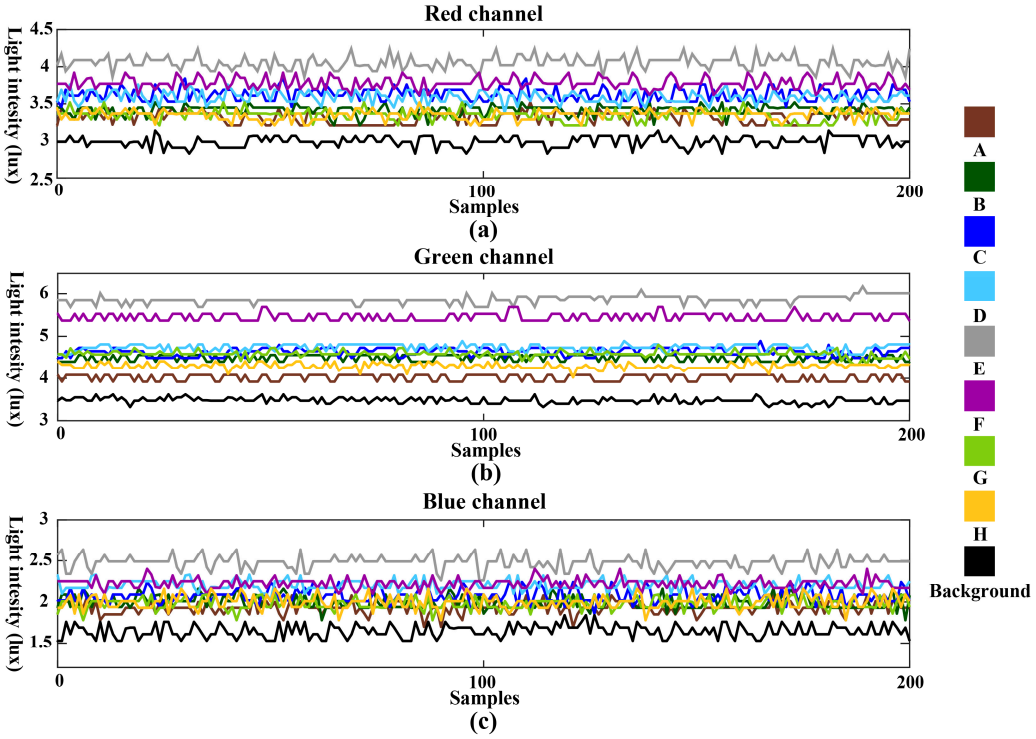

Figure 6. The light intensity data from RGB three-channel spectral receivers under dark conditions for the eight gestures and control group: (**a**) data from the red channel; (**b**) data from the green channel; (**c**) data from the blue channel.

In Dataset 2, data corresponding to eight gestures were collected under ambient light conditions. The purpose of this group of data collection is to test the influence of ambient light on gesture recognition accuracy. The ambient light source is a commonly used PWM modulated ceiling lamp, with an average light intensity of 65 lux measured by narrowband spectral receivers. The light-intensity data for the eight gestures collected from the RGB three channels, as well as the control group, are shown in Figure 7. Variations in different gestures not only affect changes in screen-reflected light but also influence the reception of ambient light by the narrowband spectral receivers. Consequently, the changes in light intensity captured by the narrowband spectral receivers integrate the

effects of both factors. The impact of gestures on ambient light comprises reflections and obstruction of light caused by the gestures, with the proportion depending on spatial positioning and spectral wavelength. For instance, in the green channel, the light intensity data for the control group without gestures are lowest, indicating a significant effect of gestures on the reflection of green light. Conversely, in the red channel, the data show the opposite trend, with the light intensity for the control group without gestures being highest, indicating a greater impact of gestures on obstructing red light. The blue channel data exhibit a more balanced effect from both factors, resulting in less discernible features visually. Therefore, in Dataset 2 as shown in Figure 7, the complex lighting conditions lead to more pronounced differences in the distribution of data across different channels. Such complex illumination environments necessitate the integration of spatial information and multi-band narrowband spectra to capture multidimensional features effectively, thereby enhancing gesture recognition accuracy.

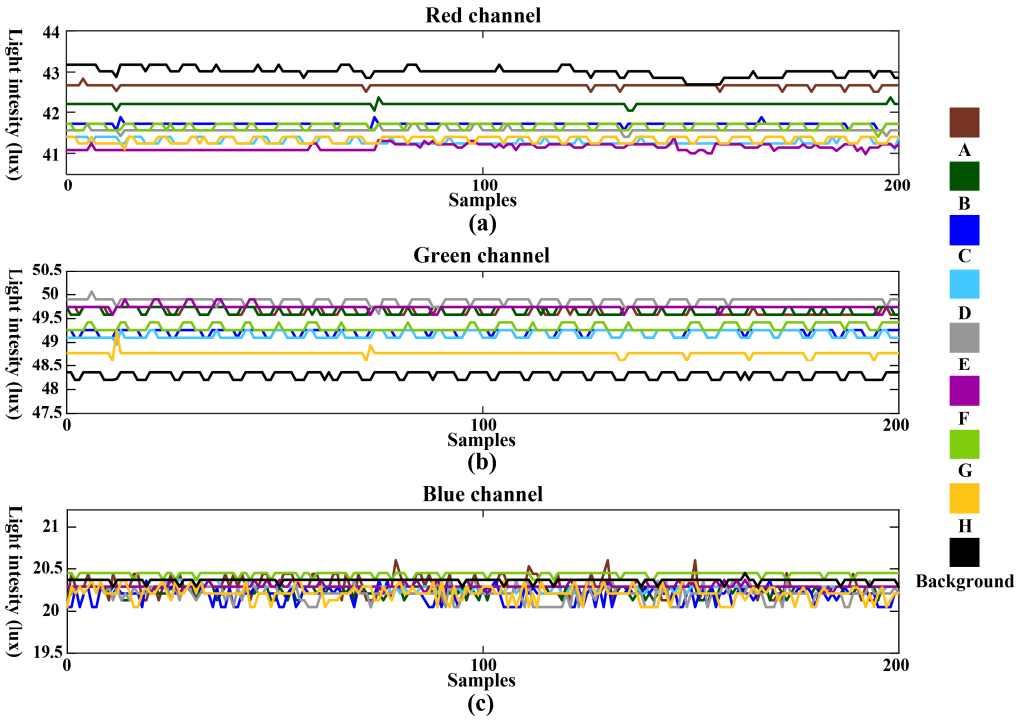

Figure 7. The light intensity data from RGB three-channel spectral receivers under ambient light conditions for the eight gestures and control group: (**a**) data from the red channel; (**b**) data from the green channel; (**c**) data from the blue channel.

During the data collection process for both datasets, each gesture sample was captured for 1 s, with each sample containing 100 samplings. The gesture data were sourced from 10 volunteers, comprising 5 men and 5 women. Variations in their hand sizes and skin tones resulted in different effects on the reflected spectrum. During collection, each volunteer's hand was positioned 30 cm from the screen, with each gesture from each individual being sampled 92 times, corresponding to 92 different images with various color tones displayed on the screen. Each dataset consisted of 920 samples per gesture, totaling 920 × 8 data points, as shown in Table 1. A randomly selected quarter of the dataset was designated as the test set.

Table 1. Description of the datasets.

Dataset	Light Source	Volunteers	Number of Gestures	Samples
1	Screen	10	8	920 × 8
2	Screen + ambient light	10	8	920 × 8

Normalization was applied to the collected data before analysis. This process transformed the time series of each sample into a one-dimensional matrix, ensuring values ranged between −1 and 1. The normalization formula is as follows:

$$x = -1 + \frac{x - x_{min}}{x_{max} - x_{min}} \times 2 \quad (1)$$

4. RGB Multi-Channel CNN-LSTM Gesture Recognition Model

This section presents a gesture recognition model that integrates RGB multi-channel 1-dimensional convolutional neural network (1D-CNN) and LSTM architectures. The model processes RGB three-channel light intensity time series as the input and generates gesture classification predictions as the output, as depicted in Figure 8. Initially, RGB multi-channel 1D-CNN is employed to extract multidimensional features from the input time-series data. Subsequently, these feature sequences are fed into LSTM for gesture classification. This hybrid approach effectively harnesses the feature extraction capabilities of 1D-CNN and the sequence modeling capabilities of LSTM. It synergizes with the multi-channel spectral information acquisition capability of the system hardware in this work, enabling accurate gesture-based HCI.

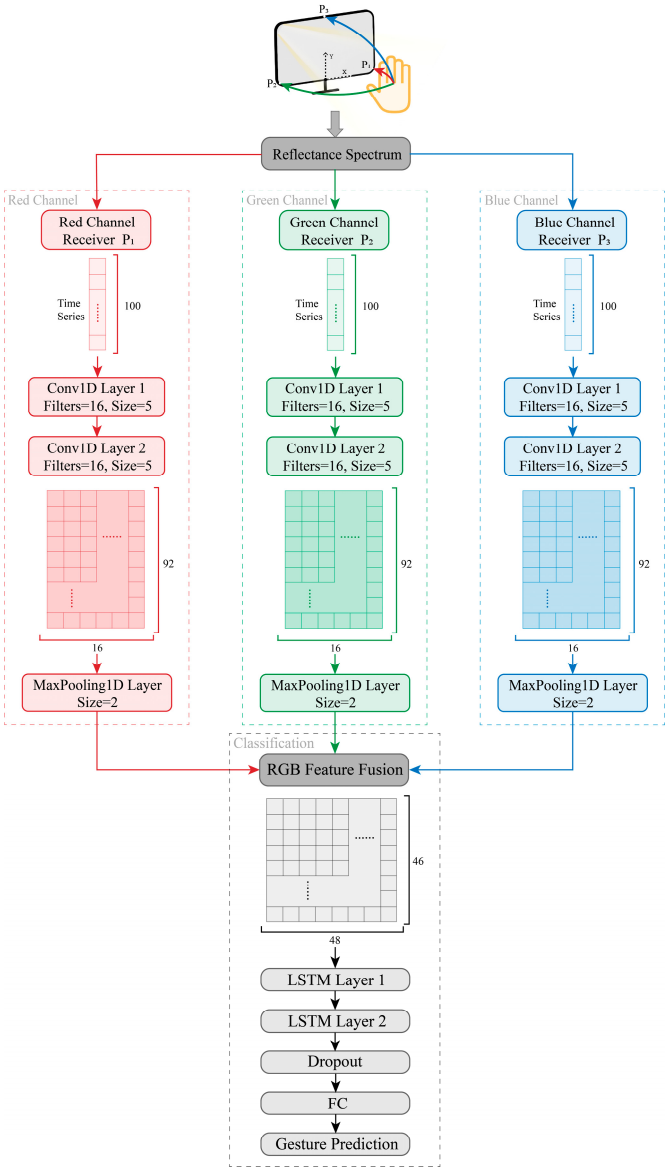

Figure 8. The RGB multi-channel CNN-LSTM gesture recognition model.

4.1. RGB Three-Channel 1D-CNN Feature Extractor

CNN can serve as a feature extractor [44], specifically, employing 1D-CNN for the analysis of time-series data from sensors. Accordingly, two layers of 1-dimensional convolution (Conv1D) are employed to extract multidimensional features from RGB three-channel data. Figure 9 illustrates the process of extracting time-series features. Each sample input to the gesture recognition model consists of 100 data points sampled over 1 s. Filters convolve with the time series to extract features. Each Conv1D layer incorporates 16 filters with a window size of 5, sliding down the data with a default stride of 1. The input data comprise RGB three channels, forming a 100 × 3 matrix that corresponds to the three channels in the Conv1D layers. Each channel shares the same structure but employs different filter combinations based on the data characteristics, thus enhancing the representation of the input time-series features. Finally, following a 1-dimensional max pooling (MaxPooling1D) layer with a size of 2, the features are merged across multiple channels as shown in Figure 8, resulting in each sample being represented as a multidimensional feature sequence of size 46 × 48 matrix.

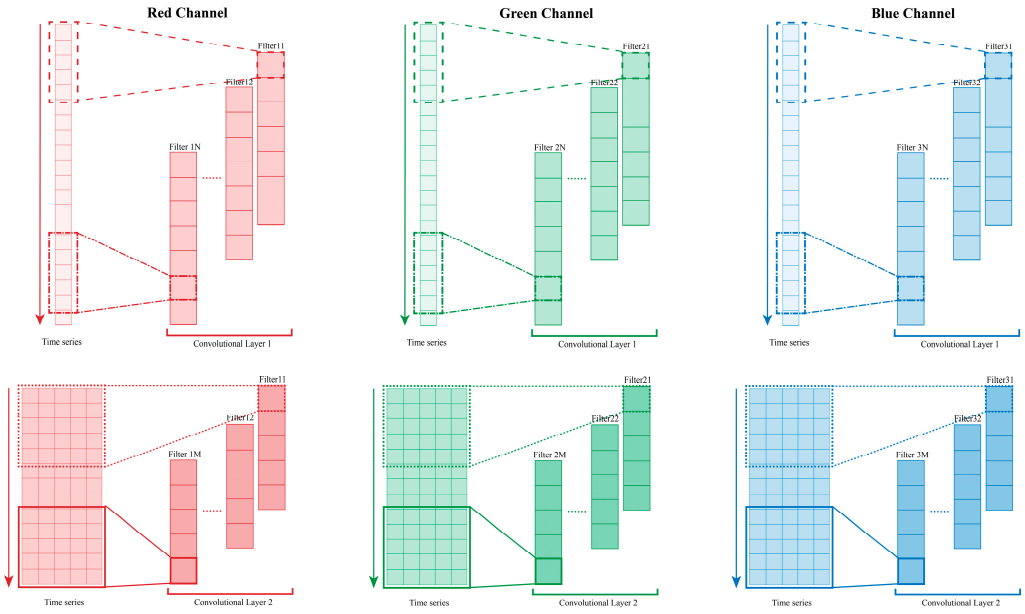

Figure 9. The feature extraction process of RGB multi-channel 1D-CNN.

4.2. LSTM Network

LSTM networks [45], a specialized category of recurrent neural networks (RNNs), are proficient in identifying and forecasting both short-term and long-term dependencies within time-series data [46]. Information is transmitted among different cells of the hidden layer through several controllable gates [47], as depicted in Figure 10. The symbol c represents the memory cell state. The network contains input gate i_t, output gate o_t, and forget gate f_t. The input gate i_t determines the contributions of the input data at time step t for updating the memory cell, while the forget gate f_t determines how much of the last moment's cell c_{t-1} is retained for the current state c_t. The output gate o_t controls how much information is output for cell status. Finally, $\widetilde{c}_t$ represents the next state. The LSTM network updates its information through the following Equations (2)–(9):

$$i_t = \sigma_i(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{2}$$

$$o_t = \sigma_o(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{3}$$

$$f_t = \sigma_f\left(W_f \cdot [h_{t-1}, x_t] + b_f\right) \tag{4}$$

$$\tilde{c}_t = tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \tag{5}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \tag{6}$$

$$h_t = o_t \odot tanh(c_t) \tag{7}$$

$$sigmoid(x) = \frac{1}{1 + e^{-x}} \tag{8}$$

$$tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \tag{9}$$

where W_i, W_o, W_f, and W_c represent the input weights; b_i, b_o, b_f, and b_c represent the bias weights; $\odot$ denotes element-wise product; σ represents the sigmoid function as Equation (8), and the hyperbolic tangent function is illustrated in Equation (9); and h_t represents the output. The classifier consists of two layers of LSTM, one dropout layer and one fully connected (FC) layer, and finally, uses softmax activation to output gesture labels. Training is conducted using the Adam optimizer with a learning rate of 0.001, a batch size of 27, and 128 nodes in the hidden layers. The model was developed and trained on the Anaconda3 platform, utilizing an NVIDIA GeForce RTX 3070 GPU.

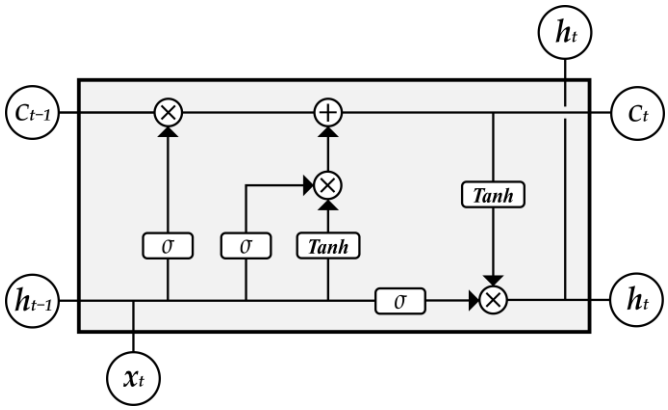

Figure 10. Network structure of LSTM.

4.3. Evaluation

In this work, macro averaging [48] is employed to evaluate the performance metrics of the multi-class classification model, including accuracy, precision, recall, and F-score [49]. These metrics are expressed by the following Formulas (10)–(13), where TP = true positives, FP = false positives, FN = false negatives, and TN = true negatives. Accuracy is the most used empirical measure, which is the ratio of the number of correct predictions to the total number of predictions.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}} \tag{10}$$

Precision is the ratio of the correct positive predictions to the total number of predictions as positives.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{11}$$

Recall is the ratio of the correct positive predictions to the total number of positive instances, also known as sensitivity.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} = \text{Sensitivity} \tag{12}$$

F-score is the harmonic mean of the precision and recall, evenly balanced when $\beta = 1$. Higher values of the F-score indicate a better balance between precision and recall.

$$\text{F-score} = \frac{(\beta^2 + 1) * \text{Precision} \times \text{Recall}}{\beta^2 * \text{Precision} + \text{Recall}} \tag{13}$$

5. Experimental Results and Discussion

The experiments on gesture recognition are divided into two steps, labeled as Experiment I and Experiment II, applied separately to Dataset 1 and Dataset 2. Each dataset comprises 5520 samples for training and 1840 samples for testing. Experiment I evaluates the performance of the gesture recognition system under dark conditions using only screen-reflected spectra. Experiment II assesses the performance in the presence of ambient light sources, considering the combined effects of screen-reflected light and external illumination.

5.1. Experimental I Results

Figure 11 presents the confusion matrix results for Experiment I evaluated on Dataset 1. Figure 11a depicts the confusion matrix for RGB three-channel gesture classification, indicating an accuracy of 99.93%, with accuracies exceeding 99% for all eight gestures. Detailed performance metrics are listed in Table 2, where the precision, recall, and F1-score of this classification model all achieve 99.73%. To demonstrate the efficacy of multi-band spectral features in enhancing gesture recognition, the classification results of Dataset 1 are compared between the RGB three-channel and single-channel. The single-channel classification employs data from either the red, green, or blue channel, based on the single-channel CNN-LSTM gesture recognition model. Figure 11b–d show the confusion matrices for the red, green, and blue channels, respectively, with accuracies of 96.45%, 95.82%, and 98.07%. Results from Table 2 indicate inferior metrics for precision, recall, and F1-score in the single-channel classification, highlighting the superior performance of the multi-channel classification model across all metrics compared with the single-channel classification models.

For a clearer comparison, Figure 12 displays the recall results for each class across the different classification models. Recall assesses the classifier's ability to correctly identify all positive instances [49]. Analysis of the curves in Figure 12 reveals varying effectiveness of different channels in recognizing each gesture. For example, the classification model trained on the red channel performs poorly for gesture G due to similarities in light intensity with gesture B, as observed in the sampling data of Figure 6, resulting in misclassification of G. Similarly, the green channel shows inadequate recognition of gesture B. In the blue channel, gestures C and D exhibit frequent confusion while demonstrating robust performance for other gestures. The disparate recognition performances across single channels highlight distinct spectral characteristics. Screen-reflected light for the same gesture exhibits spectral variation across different coordinates and is captured by diverse narrowband receivers, further differentiating the data from each channel. Additionally, features of the single channels are limited, leading to notably poorer recognition of specific gestures. In contrast, the multi-channel classification model mitigates these challenges by combining RGB three-channel spectral data from different spatial coordinates, thereby improving the accuracy of gesture recognition. In conclusion, the integration of multi-band spectral features with spatial information markedly enhances the accuracy of gesture recognition.

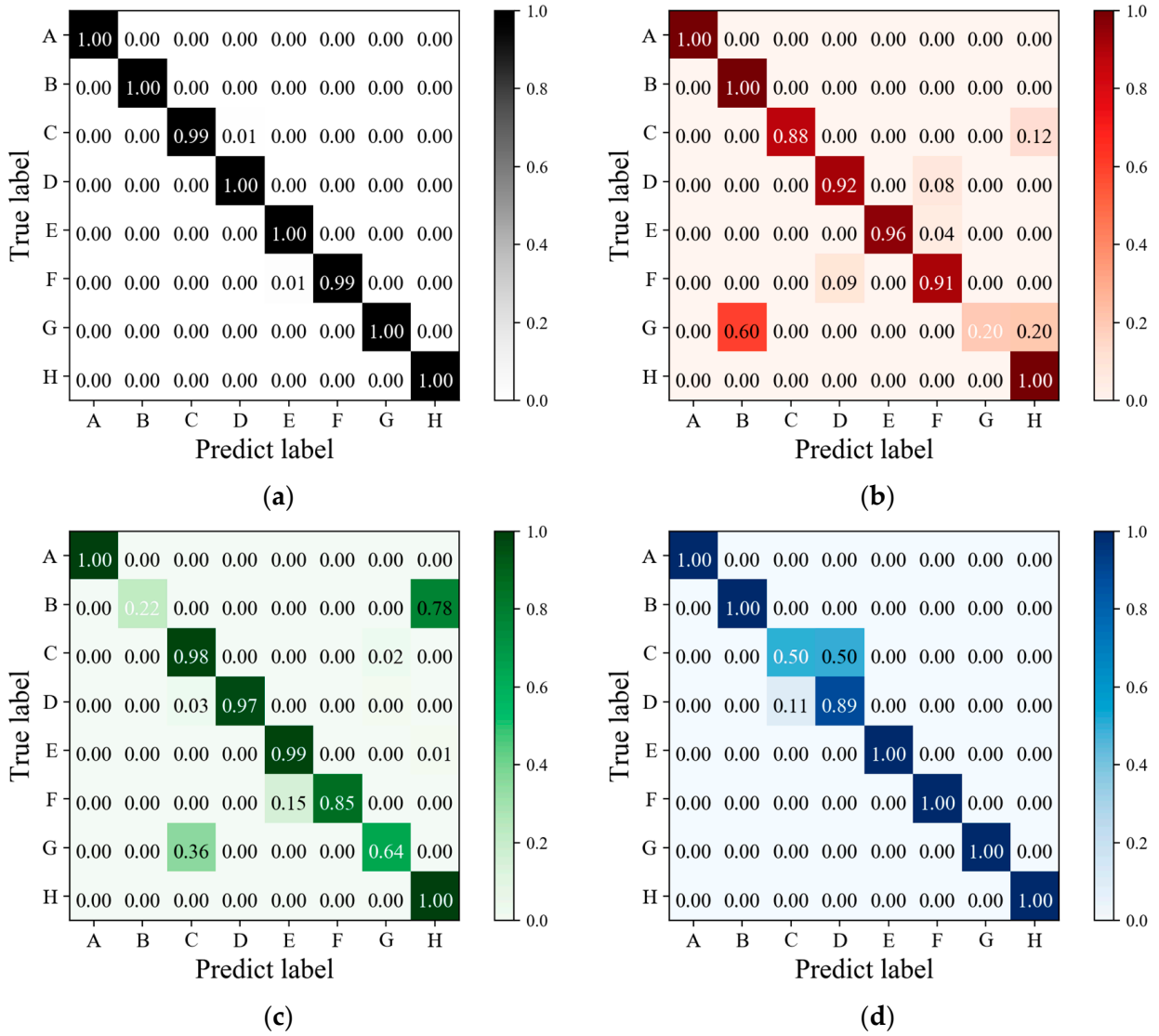

Figure 11. The confusion matrix results of Experiment I: (**a**) The confusion matrix result of RGB three-channel gesture classification; (**b**) the confusion matrix result of red-channel gesture classification; (**c**) the confusion matrix result of green-channel gesture classification; (**d**) the confusion matrix result of blue-channel gesture classification.

Table 2. Evaluation metrics of the classification models in experiments.

Experiment	Channel	Accuracy	Precision	Recall	F1-Score
I	RGB three-channel	99.93%	99.73%	99.73%	99.73%
	Red channel	96.45%	89.66%	85.82%	83.59%
	Green channel	95.82%	88.94%	83.26%	81.43%
	Blue channel	98.07%	93.15%	92.28%	91.98%
II	RGB three-channel	99.89%	99.57%	99.57%	99.57%
	Red channel	94.16%	78.53%	76.63%	74.42%
	Green channel	96.56%	85.94%	86.25%	85.32%
	Blue channel	89.29%	63.62%	57.17%	54.45%

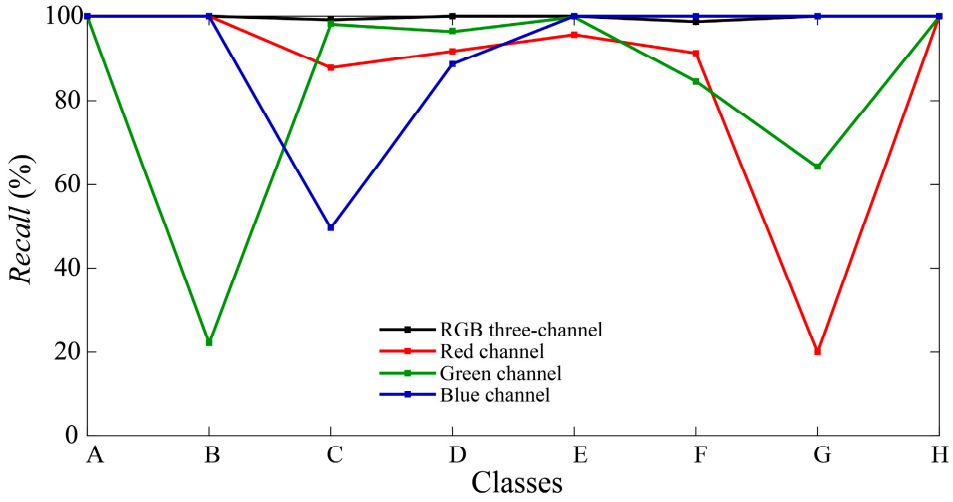

Figure 12. The recall results for each class across different classification models in Experiment I.

5.2. Experimental II Results

Figure 13 illustrates the confusion matrix results of Experiment II evaluated on Dataset 2. Metrics for all classification models are listed in Table 2. The confusion matrix for the RGB three-channel classification model is shown in Figure 13a. Despite the more complex composition of light sources in Experiment II, the classification results remain highly accurate, with an accuracy of 99.89%. The model achieves precision, recall, and F1-score metrics of 99.57%, indicating that the proposed gesture recognition method and system can operate effectively even in the presence of external light sources.

Additionally, this step of the experiment also evaluates the classification of single-channel data. Figure 13b–d depict the confusion matrices for the red, green, and blue channels, respectively, with accuracies of 94.16%, 96.56%, and 89.29%. Results from Table 2 demonstrate lower performance metrics for the single-channel classification, underscoring a substantial disparity when compared with the multi-channel classification model. Recall results for each class are compared in Figure 14, revealing that the red channel model performs poorly in recognizing gesture E, the green channel struggles with gesture C, and the overall classification performance in the blue channel is inadequate. Based on the analysis of light intensity data in Figure 7, in the presence of ambient light, narrowband spectral receivers integrate light information affected by gesture interaction with both screen light and ambient light. Different spatial positions and spectral wavelengths influence single-channel performance differently: the green channel is mainly influenced by reflected light, the red channel by shadows of ambient light caused by gestures, and the blue channel by varying light intensity due to both reflected light and shadows from gestures. Consequently, the characteristics captured by the blue channel are not sufficiently distinct, resulting in poor classification performance. In such complex lighting environments, the integration of multi-band spectral data from multiple spatial positions becomes crucial. Even in cases where individual single-channel classifications perform poorly, such as gesture C, with recall values of 87.39%, 26.52%, and 24.35% in the red, green, and blue channels, respectively, the combination of spatial and spectral features with a multi-channel gesture recognition model effectively raises the recall of gesture C to 99.13%. This synergistic effect demonstrates how the combined utilization of multiple channels yields superior performance compared with each channel individually.

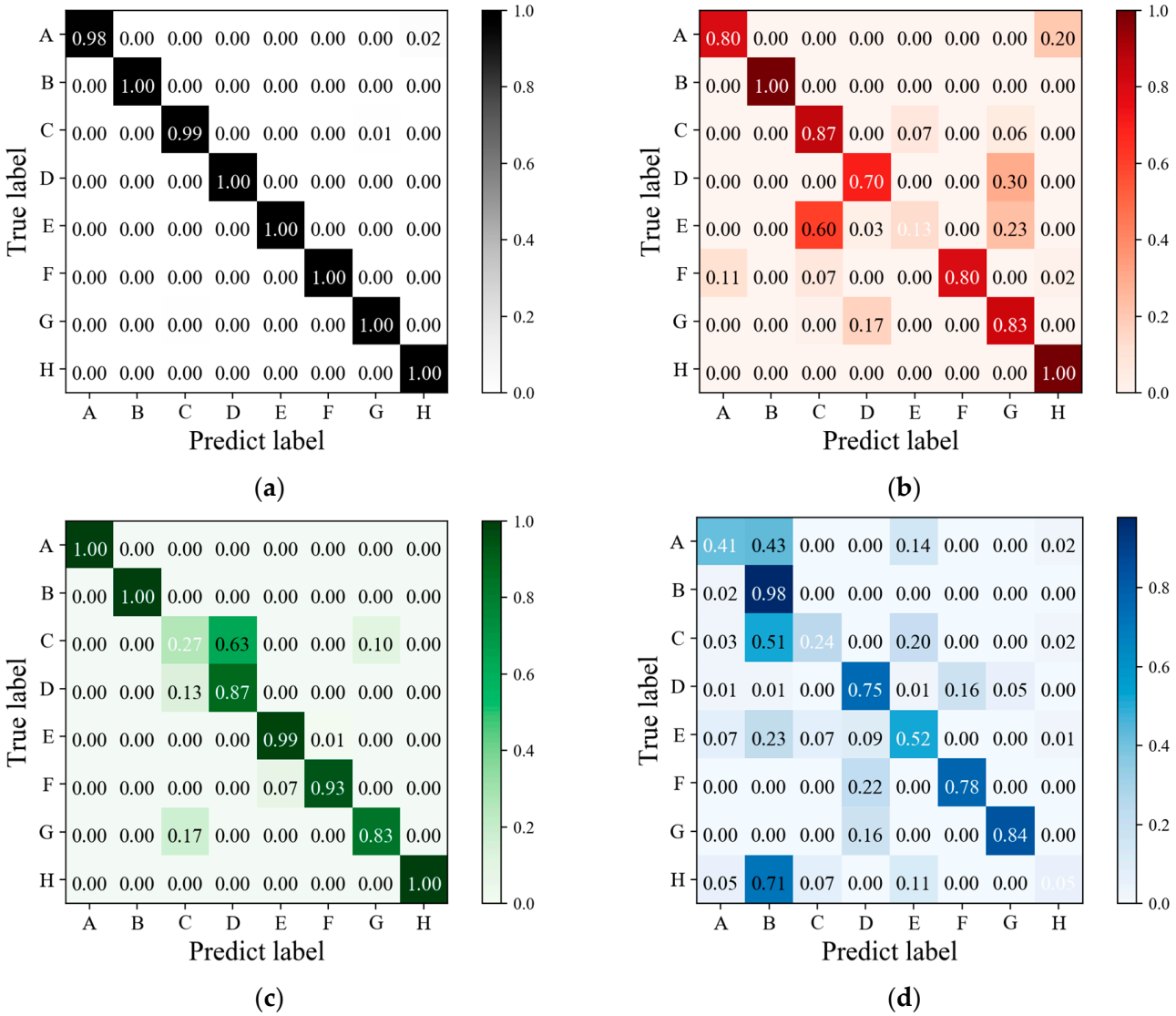

Figure 13. The confusion matrix results of experiment II: (**a**) The confusion matrix result of RGB three-channel gesture classification; (**b**) the confusion matrix result of red-channel gesture classification; (**c**) the confusion matrix result of green-channel gesture classification; (**d**) the confusion matrix result of blue-channel gesture classification.

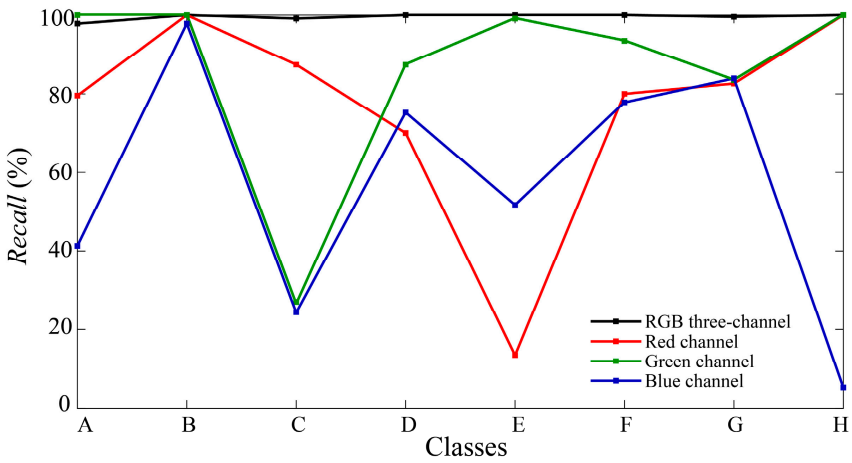

Figure 14. The recall results for each class across different classification models in Experiment II.

5.3. Discussion

Based on the results above, we summarize the experimental results and discuss future directions for improvements.

The experiments validated the proposed gesture recognition method that integrates multi-band spectral data with spatial information. Experiment I, conducted in darkness, demonstrated high accuracy of 99.93% using only screen-reflected spectra for gesture recognition. In Experiment II, which introduced ambient light sources in complex lighting environments, single-channel recognition performed poorly. In contrast, the proposed multi-band spectral gesture recognition model maintained effective performance, significantly enhancing recognition accuracy to 99.89% compared with the single-channel models. The system is well-suited for indoor applications.

We compared the experimental results with other recent non-contact screen interaction systems that employ gesture recognition, as listed in Table 3. In similar research, computer vision [19,20] is commonly used, with performance dependent on image quality and camera specifications. Passive sound sensing [21] is another convenient gesture interaction method but, like computer vision, it faces privacy and security concerns. Cheng et al. [23] developed a radar-based system with high recognition accuracy, though it incurs significant equipment costs. In the design by researchers Liao et al. [42], a single light sensor was installed on the screen, necessitating coordination with the display content and leaving room for further system improvements. Our proposed system offers advantages, including lower cost and easier portability of the narrowband spectral receivers. It addresses privacy and security concerns while achieving a relatively high level of recognition accuracy. However, the limitation lies in the restricted range of recognizable gestures and scenarios. We propose the following directions for improvements.

Table 3. Comparison with other recent non-contact screen interaction systems based on gesture recognition.

System	Equipment	Accuracy	Number of Gestures	Algorithm
Zahra et al. [19]	Camera	93.35%	6	Skin detection and genetic algorithm
Benitez-Garcia et al. [20]	Camera	85.10%	13	Temporal segment networks (TSN), temporal shift modules (TSM)
Luo et al. [21]	Microphone	93.20%	7	Feature extraction and support vector machine (SVM)
Cheng et al. [23]	Millimeter wave radar and a thermal imager	100.00%	5	Feature extraction and gated recurrent unit (GRU)
Liao et al. [42]	Ambient light sensor	96.10%	9	Feature extraction and k-nearest neighbors (KNN)
This work	Narrowband spectral receivers	99.93%	8	RGB multi-channel CNN-LSTM

(1) The gesture categories and spectral ranges in this work are limited. In future research, expanding the range of gesture classifications could enable more complex human–machine interactions, potentially incorporating dynamic movements. For example, integrating with a sign language database would greatly enhance the system's practicality for individuals with hearing and speech impairments. To achieve this, detailed plans for data collection and window segmentation will be essential. Additionally, this work focused solely on collecting spectral data within the visible light range. Future extensions could involve expanding to wider spectral ranges to fully leverage the data characteristics across different spectra.

(2) This work established an RGB three-channel narrowband spectral gesture recognition system. Future efforts will focus on optimizing the reception system to advance the

accuracy and applicability of the proposed method in diverse real-world scenarios. To enhance accuracy in complex interactions, deploying more narrowband receivers at multiple locations to establish a reception matrix would prove beneficial.

6. Conclusions

The intelligent gesture recognition method proposed in this paper leverages multi-band spectral features that integrate frequency domain and spatial domain information to enhance accuracy. Based on this method, an RGB three-channel narrowband spectral gesture recognition system is developed, incorporating a screen and multiple narrowband spectral receivers as essential hardware components. Integrated with the RGB multi-channel CNN-LSTM classification model, the system accurately recognizes eight types of gestures and enables interaction with display screens. It processes multi-channel time series data from narrowband spectral receivers, achieving accuracies of 99.93% in dark conditions and 99.89% in illuminated conditions. The collaborative effect of the multi-channel features enhances performance, significantly improving recognition accuracy compared with single-channel models. This gesture recognition method offers straightforward implementation, ensuring privacy and security and facilitating its widespread application in various screen-based human–machine interactions.

Author Contributions: Conceptualization, P.L. and J.H.; methodology, P.L.; formal analysis, P.L. and C.L.; investigation, P.L. and W.Y.; data curation, P.L. and S.C.; writing—original draft preparation, P.L.; writing—review and editing, P.L.; supervision, J.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are contained within the article.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Vrana, J.; Singh, R. *Handbook of Nondestructive Evaluation 4.0*; Springer International Publishing: Berlin/Heidelberg, Germany, 2022; pp. 107–123.
2. Hewett, T.; Baecker, R.; Card, S.; Carey, T.; Gasen, J.; Mantei, M.; Perlman, G.; Strong, G.; Verplank, W. *ACM SIGCHI Curricula for Human-Computer Interaction*; ACM Press: New York, NY, USA, 1992; pp. 5–7.
3. Mourtzis, D.; Angelopoulos, J.; Panopoulos, N. The future of the human–machine interface (HMI) in society 5.0. *Future Internet* **2023**, *15*, 162. [CrossRef]
4. Reipschlager, P.; Flemisch, T.; Dachselt, R. Personal augmented reality for information visualization on large interactive displays. *IEEE Trans. Vis. Comput. Graph.* **2021**, *27*, 1182–1192. [CrossRef] [PubMed]
5. Biele, C. Hand movements using keyboard and mouse. *Hum. Mov. Hum.-Comput. Interact.* **2022**, *996*, 39–51.
6. Wu, J.; Zhu, Y.; Fang, X.; Banerjee, P. Touch or click? The effect of direct and indirect human-computer interaction on consumer responses. *J. Mark. Theory Pract.* **2023**, *32*, 158–173. [CrossRef]
7. Jakobsen, M.R.; Hornbaek, K. Up close and personal: Collaborative work on a high-resolution multitouch wall display. *ACM Trans. Comput.-Hum. Interact.* **2014**, *21*, 1–34. [CrossRef]
8. Nunes, J.S.; Castro, N.; Gonçalves, S.; Pereira, N.; Correia, V.; Lanceros-Mendez, S. Marked object recognition multitouch screen printed touchpad for interactive applications. *Sensors* **2017**, *17*, 2786. [CrossRef]
9. Prouzeau, A.; Bezerianos, A.; Chapuis, O. Evaluating multi-user selection for exploring graph topology on wall-displays. *IEEE Trans. Vis. Comput. Graph.* **2016**, *23*, 1936–1951. [CrossRef] [PubMed]
10. Huang, Z.; Huang, X. A study on the application of voice interaction in automotive human machine interface experience design. In Proceedings of the AIP Conference, Xi'an, China, 20–21 January 2018; p. 040074.
11. Uludağli, M.Ç.; Acartürk, C. User interaction in hands-free gaming: A comparative study of gaze-voice and touchscreen interface control. *Turk. J. Electr. Eng. Comput. Sci.* **2018**, *26*, 1967–1976. [CrossRef]
12. Gao, L.; Liu, Y.; Le, J.; Liu, R. Research on the application of multi-channel interaction in information system. In Proceedings of the 2nd International Conference on Robotics, Artificial Intelligence and Intelligent Control (RAIIC), Mianyang, China, 11–13 August 2023; pp. 121–125.

13. Birch, B.; Griffiths, C.A.; Morgan, A. Environmental effects on reliability and accuracy of MFCC based voice recognition for industrial human-robot-interaction. *Proc. Inst. Mech. Eng. B J. Eng. Manuf.* **2021**, *235*, 1939–1948. [CrossRef]
14. Alrowais, F.; Negm, N.; Khalid, M.; Almalki, N.; Marzouk, R.; Mohamed, A.; Al Duhayyim, M.; Alneil, A.A. Modified earthworm optimization with deep learning assisted emotion recognition for human computer interface. *IEEE Access* **2023**, *11*, 35089–35096. [CrossRef]
15. Pereira, R.; Mendes, C.; Ribeiro, J.; Ribeiro, R.; Miragaia, R.; Rodrigues, N.; Costa, N.; Pereira, A. Systematic review of emotion detection with computer vision and deep learning. *Sensors* **2024**, *24*, 3484. [CrossRef] [PubMed]
16. Aghajanzadeh, S.; Naidu, R.; Chen, S.H.; Tung, C.; Goel, A.; Lu, Y.H.; Thiruvathukal, G.K. Camera placement meeting restrictions of computer vision. In Proceedings of the IEEE International Conference on Image Processing (ICIP), Abu Dhabi, United Arab Emirates, 25–28 October 2020; pp. 3254–3258.
17. Harshitaa, A.; Hansini, P.; Asha, P. Gesture based home appliance control system for disabled people. In Proceedings of the Second International Conference on Electronics and Sustainable Communication Systems (ICESC), Coimbatore, India, 4–6 August 2021; pp. 1501–1505.
18. Ryumin, D.; Ivanko, D.; Axyonov, A. Cross-language transfer learning using visual information for automatic sign gesture recognition. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.* **2023**, *48*, 209–216. [CrossRef]
19. Zahra, R.; Shehzadi, A.; Sharif, M.I.; Karim, A.; Azam, S.; De Boer, F.; Jonkman, M.; Mehmood, M. Camera-based interactive wall display using hand gesture recognition. *Intell. Syst. Appl.* **2023**, *19*, 200262. [CrossRef]
20. Benitez-Garcia, G.; Prudente-Tixteco, L.; Castro-Madrid, L.C.; Toscano-Medina, R.; Olivares-Mercado, J.; Sanchez-Perez, G.; Villalba, L.J.G. Improving real-time hand gesture recognition with semantic segmentation. *Sensors* **2021**, *21*, 356. [CrossRef]
21. Luo, G.; Yang, P.; Chen, M.; Li, P. HCI on the table: Robust gesture recognition using acoustic sensing in your hand. *IEEE Access* **2020**, *8*, 31481–31498. [CrossRef]
22. Hazra, S.; Santra, A. Robust gesture recognition using millimetric-wave radar system. *IEEE Sens. Lett.* **2018**, *2*, 1–4. [CrossRef]
23. Cheng, Y.L.; Yeh, W.; Liao, Y.P. The implementation of a gesture recognition system with a millimeter wave and thermal imager. *Sensors* **2024**, *24*, 581. [CrossRef] [PubMed]
24. Oudah, M.; Al-Naji, A.; Chahl, J. Hand gesture recognition based on computer vision: A review of techniques. *J. Imaging* **2020**, *6*, 73. [CrossRef]
25. Galván-Ruiz, J.; Travieso-González, C.M.; Tejera-Fettmilch, A.; Pinan-Roescher, A.; Esteban-Hernández, L.; Domínguez-Quintana, L. Perspective and evolution of gesture recognition for sign language: A review. *Sensors* **2020**, *20*, 3571. [CrossRef]
26. Sokhib, T.; Whangbo, T.K. A combined method of skin-and depth-based hand gesture recognition. *Int. Arab J. Inf. Technol.* **2020**, *17*, 137–145. [CrossRef]
27. Xu, J.; Li, J.; Zhang, S.; Xie, C.; Dong, J. Skeleton guided conflict-free hand gesture recognition for robot control. In Proceedings of the 11th International Conference on Awareness Science and Technology (iCAST), Qingdao, China, 7–9 December 2020; pp. 1–6.
28. Alwaely, B.; Abhayaratne, C. Ghosm: Graph-based hybrid outline and skeleton modelling for shape recognition. *ACM Trans. Multim. Comput. Commun. Appl.* **2023**, *19*, 1–23. [CrossRef]
29. Qiao, G.; Ning, N.; Zuo, Y.; Zhou, P.; Sun, M.; Hu, S.; Yu, Q.; Liu, Y. Spatio-temporal fusion spiking neural network for frame-based and event-based camera sensor fusion. *IEEE Trans. Emerg. Top. Comput. Intell.* **2024**, *8*, 2446–2456. [CrossRef]
30. Ryumin, D.; Ivanko, D.; Ryumina, E. Audio-visual speech and gesture recognition by sensors of mobile devices. *Sensors* **2023**, *23*, 2284. [CrossRef] [PubMed]
31. Hakim, N.L.; Shih, T.K.; Kasthuri Arachchi, S.P.; Aditya, W.; Chen, Y.C.; Lin, C.Y. Dynamic hand gesture recognition using 3DCNN and LSTM with FSM context-aware model. *Sensors* **2019**, *19*, 5429. [CrossRef] [PubMed]
32. Sharma, P.; Anand, R.S. Depth data and fusion of feature descriptors for static gesture recognition. *IET Image Process.* **2020**, *14*, 909–920. [CrossRef]
33. Zengeler, N.; Kopinski, T.; Handmann, U. Hand gesture recognition in automotive human–machine interaction using depth cameras. *Sensors* **2019**, *19*, 59. [CrossRef]
34. Yu, J.; Qin, M.; Zhou, S. Dynamic gesture recognition based on 2D convolutional neural network and feature fusion. *Sci. Rep.* **2022**, *12*, 4345. [CrossRef]
35. Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; Paluri, M. Learning spatiotemporal features with 3d convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 4489–4497.
36. Hui, W.S.; Huang, W.; Hu, J.; Tao, K.; Peng, Y. A new precise contactless medical image multimodal interaction system for surgical practice. *IEEE Access* **2020**, *8*, 121811–121820. [CrossRef]
37. Safavi, S.M.; Sundaram, S.M.; Heydarigorji, A.; Udaiwal, N.S.; Chou, P.H. Application of infrared scanning of the neck muscles to control a cursor in human-computer interface. In Proceedings of the 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Jeju, Republic of Korea, 11–15 July 2017; pp. 787–790.
38. Singh, J.; Raza, U. Passive visible light positioning systems: An overview. In Proceedings of the Workshop on Light Up the IoT, London, UK, 21 September 2020; pp. 48–53.
39. Fragner, C.; Krutzler, C.; Weiss, A.P.; Leitgeb, E. LEDPOS: Indoor visible light positioning based on LED as sensor and machine learning. *IEEE Access* **2024**, *12*, 46444–46461. [CrossRef]

40. Pathak, P.H.; Feng, X.; Hu, P.; Mohapatra, P. Visible light communication, networking, and sensing: A survey, potential and challenges. *IEEE Commun. Surv. Tutor.* **2015**, *17*, 2047–2077. [CrossRef]
41. Lu, Y.; Wu, F.; Huang, Q.; Tang, S.; Chen, G. Telling secrets in the light: An efficient key extraction mechanism via ambient light. *IEEE Trans. Wirel. Commun.* **2021**, *20*, 186–198. [CrossRef]
42. Liao, Z.; Luo, Z.; Huang, Q.; Zhang, L.; Wu, F.; Zhang, Q.; Wang, Y. SMART: Screen-based gesture recognition on commodity mobile devices. In Proceedings of the 27th Annual International Conference on Mobile Computing and Networking, New Orleans, LA, USA, 31 January–4 February 2022; pp. 283–295.
43. Lin, P.; Zhuo, R.; Wang, S.; Wu, Z.; Huangfu, J. LED screen-based intelligent hand gesture recognition system. *IEEE Sens. J.* **2022**, *22*, 24439–24448. [CrossRef]
44. Jogin, M.; Madhulika, M.S.; Divya, G.D.; Meghana, R.K.; Apoorva, S. Feature extraction using convolution neural networks (CNN) and deep learning. In Proceedings of the 3rd IEEE International Conference on Recent Trends in Electronics, Information and Communication Technology (RTEICT), Bangalore, India, 18–19 May 2018; pp. 2319–2323.
45. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]
46. Sherstinsky, A. Fundamentals of recurrent neural network and long short-term memory network. *Phys. D Nonlinear Phenom.* **2020**, *404*, 132306. [CrossRef]
47. Gers, F.A.; Schmidhuber, J.; Cummins, F. Learning to forget: Continual prediction with LSTM. *Neural Comput.* **2000**, *12*, 2451–2471. [CrossRef]
48. Takahashi, K.; Yamamoto, K.; Kuchiba, A.; Koyama, T. Confidence interval for micro-averaged F_1 and macro-averaged F_1 scores. *Appl. Intell.* **2022**, *52*, 4961–4972. [CrossRef]
49. Sokolova, M.; Japkowicz, N.; Szpakowicz, S. Beyond accuracy, F-score and ROC: A family of discriminant measures for performance evaluation. In Proceedings of the 19th Australasian Joint Conference on Artificial Intelligence, Berlin, Germany, 4–8 December 2006; pp. 1015–1021.