applied sciences

MDPI

Article

Hand Gesture Detection and Recognition Using Spectrogram and Image Processing Technique with a Single Pair of Ultrasonic Transducers

Guo-Hua Feng [1],* and Gui-Rong Lai [2]

[1] Department of Power Mechanical Engineering and NEMS Institute, National Tsing Hua University, Hsinchu 300, Taiwan
[2] Department of Mechanical Engineering, National Chung Cheng University, Chiayi 621, Taiwan; n660665@yahoo.com
* Correspondence: ghfeng@pme.nthu.edu.tw

Abstract: This paper presents an effective signal processing scheme of hand gesture recognition with a superior accuracy rate of judging identical and dissimilar hand gestures. This scheme is implemented with the air sonar possessing a pair of cost-effective ultrasonic emitter and receiver along with signal processing circuitry. Through the circuitry, the Doppler signals of hand gestures are obtained and processed with the developed algorithm for recognition. Four different hand gestures of push motion, wrist motion from flexion to extension, pinch out, and hand rotation are investigated. To judge the starting time of hand gesture occurrence, the technique based on continuous short-period analysis is proposed. It could identify the starting time of the hand gesture with small-scale motion and avoid faulty judgment while no hand in front of the sonar. Fusing the short-time Fourier transform spectrogram of hand gesture to the image processing techniques of corner feature detection, feature descriptors, and Hamming-distance matching are the first-time, to our knowledge, employed to recognize hand gestures. The results show that the number of matching points is an effective parameter for classifying hand gestures. Based on the experimental data, the proposed scheme could achieve an accuracy rate of 99.8% for the hand gesture recognition.

Keywords: ultrasound; hand gesture; image processing; short-time Fourier transform; air sonar

check for updates

Citation: Feng, G.-H.; Lai, G.-R. Hand Gesture Detection and Recognition Using Spectrogram and Image Processing Technique with a Single Pair of Ultrasonic Transducers. *Appl. Sci.* **2021**, *11*, 5407. https://doi.org/10.3390/app11125407

Academic Editor: Theodore E. Matikas

Received: 22 May 2021
Accepted: 8 June 2021
Published: 10 June 2021

1. Introduction

Nowadays, ultrasound is broadly applied in several fields, and the major areas include medical and structural diagnoses and range-finding-related applications. For example, Doppler radar can be used in aviation, meteorology, speed gun, motion detection, intruder warning, and light control. In particular, Doppler radar has been used for posture and gesture recognition in motion sensing [1–5].

Posture or gesture recognition commonly uses mechanical or light stimulation sensors, such as cameras, infrared sensors, and wearable devices. Although their installation may not be difficult and the obtained results could be intuitively analyzed, these devices have inherent limitations [6]. Cameras can infringe on users' privacy; infrared sensors are susceptible to misjudgment; securing wearable sensors is necessary for functional operation [7]. In contrast, acoustic stimulus sensing, such as sonar, possesses several unique advantages, particularly for the monitoring of the home environment and can be fully operated in the dark [8]. In addition, visuals need not be recorded so that the privacy of users will not be affected [9]. These characteristics make sonar a superior sensing system for gesture recognition applications.

The hand gesture detection system possesses several advantages. The contactless interaction mode allows the users not to touch the control panel. This avoids the potential cross-contamination of multiple users via the touch panel/screen and the likely damage/fatigue of the physical control device due to inappropriate/intensified operation. In addition, the proposed hand gestures in the study are aimed at robot manipulation. The hand gestures intuitively correspond to the motion of the robot. This facilitates the users not only to operate the robot but to collaborate with the robot to execute a task.

Liu et al. used a single speaker and microphone in a smartphone with the Doppler effect to study gesture recognition via ultrasound. The vectors containing +1 and −1 can be used to indicate the direction of gesture movement in each time period [10]. The SoundWave team also used a microphone and speaker in a mobile device. The gestures investigated by them consisted of scrolling, single or double tapping, forward and backward hand gestures, and continuous motion. These gestures can be judged from the amplitude of the reflected signal [11]. Przybyla et al. developed a three-dimensional range-finding function chip. Its searchable range is up to 1 m in a 45° field of view. Low power consumption is the main feature of the chip [12]. Zhou et al. reported the ultrasonic hand gesture recognition system which is capable of recognizing 3D micro-hand gestures with a high cross-user recognition rate. This system possesses the potential to develop a practical low-power human machine interaction system [13].

On the other hand, researchers working in the field of computer vision have broadly employed image processing methods such as feature detection, description, and matching for object recognition, image classification, and retrieval. Features, such as points, edges, or corners, can be considered as the information of interest within the investigated image [14]. A good feature detector can find features possessing strong information changes that are repeatedly recognized among several images taken with different viewing angles and lighting conditions [15]. A feature descriptor is an algorithm used to process the image feature points into meaningful feature vectors. Feature descriptors encode interesting image features into a series of numbers and act as a type of numerical identifier that can be employed to differentiate one feature from another.

Researchers have conducted a considerable number of studies on image feature extraction and description and proposed several classic feature description and extraction methods, such as scale invariant feature transform (SIFT), speeded up robust features (SURF), features from accelerated segment test (FAST), and binary robust independent elementary feature (BRIEF) [16–19]. These methods obtain image feature points and their descriptors by finding the local extremum in the image and describing the features using the luminance information of their neighborhood.

Feature matching is obtained by calculating the distances between feature points in different images. For example, the SIFT feature descriptor uses the Euclidean distance as the judgment standard between descriptors, whereas the BRIEF descriptor [20] is a type of binary descriptor that uses the Hamming distance as the judgment standard to describe the correspondence between two feature points [21,22].

As described above, existing air sonar applied to hand gesture recognition by time-domain reflected signal by decoding its signal level or amplitude, the likely environmental noise and the distance between the users and speaker/microphone could decrease the accuracy rate of hand gesture recognition. Moreover, most of the image recognition techniques deals with the images directly obtained via camera. The large amount data points of images accompanying with the frame change not only requires the higher cost hardware but more computational effort compared to our proposed air sonar approach. In this study, we constructed an air sonar with a pair of ultrasonic emitters and receiver to investigate the hand gesture recognition. Through the cost-effective circuity, the acquired Doppler signals were processed for the study of hand gesture recognition. Two algorithms were developed with the one to judge the starting time of hand gestures by continuous short-time Fourier transform and the other to recognize the hand gesture by image processing with spectrogram. The Superior recognition results were obtained using the proposed scheme. Further details are described below.

2. Linear and Rotational Doppler Effect on Hand Gesture Motion

The common hand gestures include translational and rotational motion of the hand. The signals from air sonar with linear and rotational Doppler effect will be applied to our hand gesture recognition. Consider the constructed ultrasonic transmitter and receiver pair as an air sonar system fixed and located at the origin Q of the radar coordinate system (U, V, W) (Figure 1). The hand is described in the local coordinate (x, y, z) and has a translation and rotation with respect to the radar coordinate. A reference coordinate system (X, Y, Z) is introduced, which has the same translation as coordinate (x, y, z) but has no rotation with respect to the (U, V, W) coordinate.

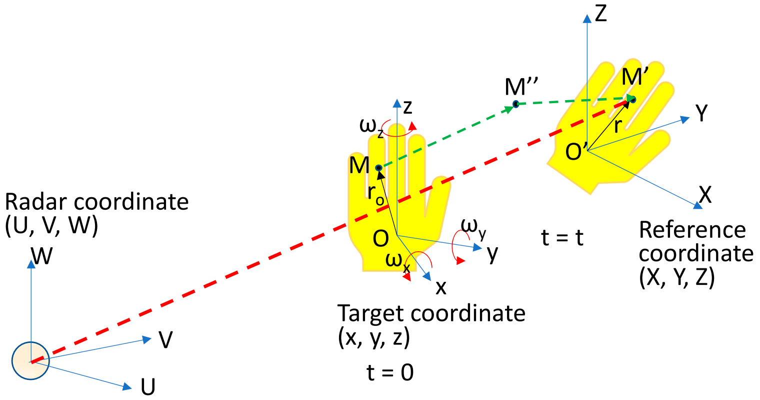
Z
M''
M'
Y
r
O'
X
Reference coordinate (X, Y, Z)
t = t
ω_z
z
M
r_o
O
ω_y
y
ω_x
x
Radar coordinate (U, V, W)
W
V
U
Target coordinate (x, y, z)
t = 0

Figure 1. The coordinate of an air sonar and a target for micro Doppler analysis.

Suppose the hand has a translation velocity v with respect to the radar and an angular rotation velocity ω represented in the reference coordinate as $\omega = (\omega_X, \omega_Y, \omega_Z)^T$. A point scatterer M on the hand, located at $r_0 = (X_0, Y_0, Z_0)^T$, at time $t = 0$ will move to M$'$ at time t. This movement can be considered as a translation from M to M$''$ with velocity v and a rotation from M$''$ to M$'$ with an angular velocity ω. At time t, the range vector from the radar to the scatterer at M$'$ becomes

$$\vec{\mathrm{OM}'} = \vec{\mathrm{QO}} + \vec{\mathrm{OO}'} + \vec{\mathrm{O'M}'} = \vec{R_0} + \vec{v}\,t + \vec{r}\,.$$

The scalar range of $\vec{\mathrm{OM}'}$ can be expressed as

$$R_t = \| \vec{R_0} + \vec{v}\,t + \vec{r} \, \| \, .$$

When the air sonar emits a continuous sinusoidal wave with a carrier frequency f_0, the echo signal $s(t)$ received from the scatterer on the hand at the position (x, y, z) is expressed as a function of R_t.

$$s(t) = \rho(x,y,z)\exp\left\{ j\left(2\pi f\frac{2R_t}{c}\right)\right\}, \tag{1}$$

where $\rho(x,y,z)$ is the reflectivity function of the point scatterer M described in the target local coordinates (x, y, z), c is the speed of sound, and the phase of the signal is $2\pi f(2R_t/c)$. The Doppler frequency shift by hand motion can be obtained by taking the time derivative of the phase as [23]

$$f_D = \frac{1}{2\pi}\frac{d(2\pi f(\frac{2R_t}{c}))}{dt} = \frac{2f}{c}\frac{dR_t}{dt} = \frac{2f}{c}\left(\vec{v} + \vec{\omega} \times \vec{r}\,\right)^T \vec{n_m}, \tag{2}$$

where $\vec{n_m} = (\vec{R_0} + \vec{v}\,t + \vec{r}\,)/R_t$.

Thus, the Doppler frequency shift included the effect of translation and rotation as well as the direction from the radar to the scatterer on the hand. Four hand gestures, namely, "push", "wrist motion from flexion to extension", "pinch out", and "hand rotation", were performed with the right hand for this study.

(1) Push: As shown in Figure 2a, the push gesture involves only a translational motion of the palm. While the hand moved forward from an initial position to a stop position, the normal direction of the palm was toward the radar and the palm underwent the acceleration, near-constant velocity, and deceleration stages. The maximum Doppler frequency shift was contributed by the center of the palm because the velocity direction was normal to the direction of the radar to the center of the palm. The minimum Doppler frequency shift resulted from the corner of the moving palm.

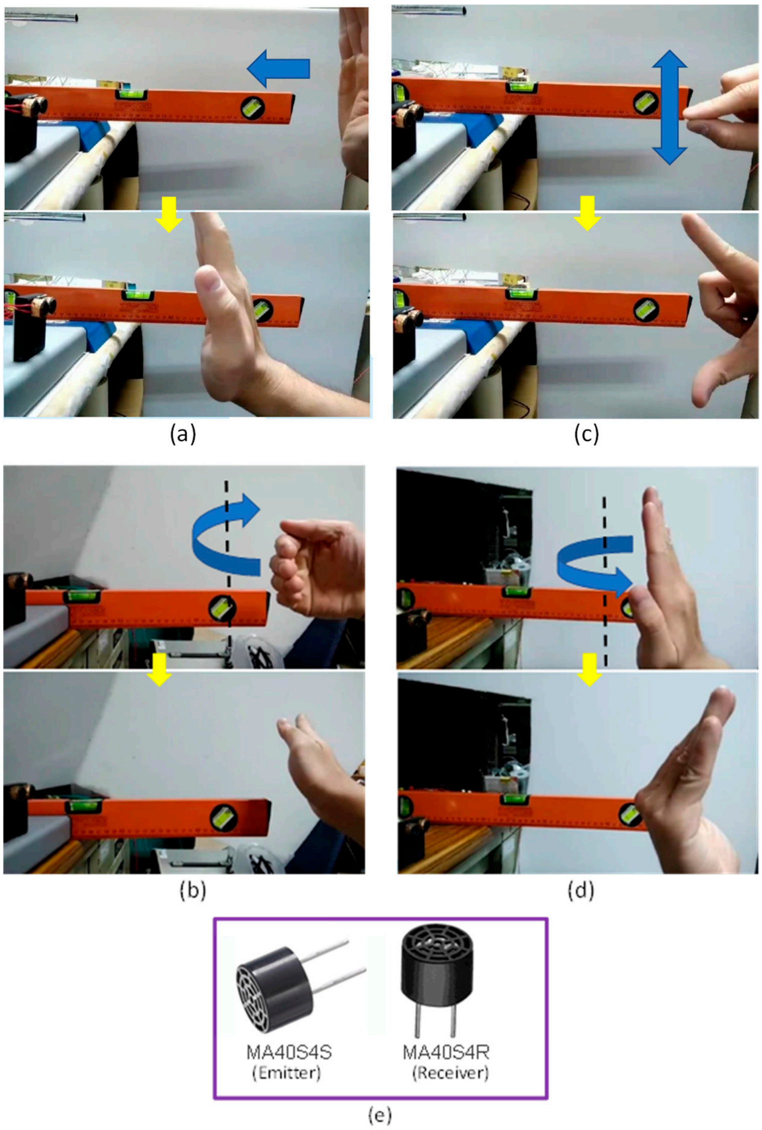

Figure 2. Hand gestures: (**a**) Push motion. (**b**) Wrist from flexion to extension. (**c**) Pinch out motion. (**d**) Hand rotation motion. (**e**) Used ultrasonic emitter and receiver.

(2) Wrist motion from flexion to extension: As shown in Figure 2b, the hand moved from the position with the palm away from the radar to that with the palm toward the radar. Initially, the wrist was flexed. Then, regarding the angular displacement, the hand underwent acceleration, constant speed, and then deceleration stages. Finally, the wrist was extended. The pivot region, which can be considered as the wrist joint, possesses

zero velocity during the entire motion, thus resulting in a zero Doppler frequency shift. The largest Doppler frequency shift was contributed by the largest value of the hand velocity at a certain scatterer point on the hand multiplying the unit direction vector from that scatterer to the radar.

(3) Pinch out: This motion is mainly performed by the thumb and index finger. In the initial state, the front ends of the index finger and thumb contacted each other. The remaining fingers were clenched. The motion is illustrated in Figure 2c. The front ends of the two fingers gradually opened up along with the forward movement of the hand. The thumb and index finger underwent acceleration from the initial position and then decelerated to a stop position, thus completing this motion. The index finger rotated in a clockwise direction, and the thumb rotated in a counterclockwise direction. These two motions provide a negative Doppler frequency shift with the magnitude gradually increasing, reaching the maximum, and then decreasing to zero. Meanwhile, the remaining fingers formed a partial fist and underwent acceleration and deceleration phases. This caused a similar Doppler frequency shift as the push motion, but covered a smaller scattering area compared with the push motion.

(4) Hand rotation: The motion of hand rotation is similar to the wrist motion from flexion to extension, with the major difference being the axis of rotation. Regarding hand rotation, the hand was aligned with the lower arm, and all the fingers closed to form a plane with the palm (Figure 2d). The axis of rotation was the centerline of the lower arm. The rotation motion started from the palm toward the radar and rotated to make the palm face away from the radar. The hand motion underwent two phases of acceleration and deceleration. The axis of rotation had zero velocity, and thus contributed zero frequency shift. During the motion, the part of the hand from the axis of rotation to the little finger edge provided a positive Doppler frequency shift, followed by a negative Doppler frequency shift. Meanwhile, the rest of the hand produced a negative Doppler frequency shift, followed by a small positive Doppler frequency shift.

Hand gesture can be served as an important tool for human interaction. Compared to existing interfaces, hand gestures have the advantages of being intuitive and easy to use. The investigated four different kinds of hand gestures which could be applied to human-robot interaction. The importance of the characterization of these hand gestures are as following: the push motion commands the robot arm to move forward, the wrist motion from flexion to extension asks the robot grip to make a right turn, the pinch out motion instructs the gripper of robot to open, and the hand rotation motion requests the robot body/platform to turn left (Figure 3).

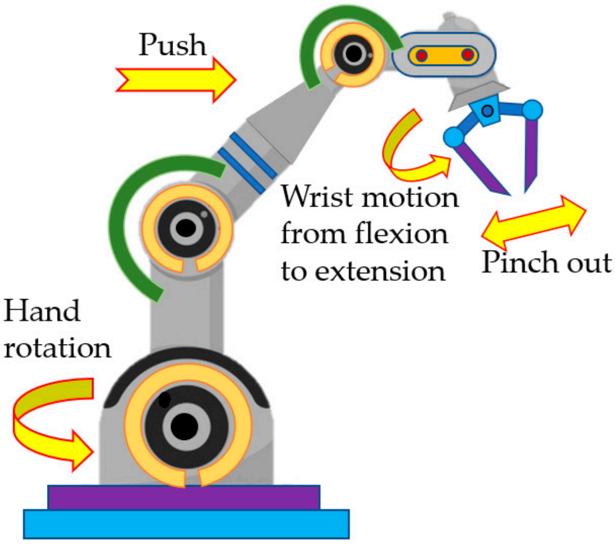

Figure 3. The potential robot operating application of the proposed hand gesture recognition.

3. Hardware Setup of Air Sonar

3.1. Ultrasonic Emitter and Receiver

We employed MA40S4S as the emitter and MA40S4R as the receiver (Murata Co., Nagaokakyo, Japan) with an operating frequency of approximately 40 kHz in the air sonar system. The sinusoidal signal was sent using a function generator to actuate the ultrasonic emitter, and an NI-USB6351 data acquisition card was used to acquire the electrical signal from the ultrasonic receiver. To avoid the sinusoidal wave generated due to distortion, we set the sampling frequency to 250 kHz, which is five times higher than the operating frequency, to examine the performance of the selected pair of ultrasonic transducers.

Two tests were performed under a sinusoidal voltage input of 20 Vpp: (1) the emitter was fixed, and the receiver was placed 3, 18, and 38 cm away such that it was facing the emitter. The amplitudes of the acquired sinusoidal signals were 6.52, 0.736, and 0.188 Vpp, respectively. (2) The emitter and receiver were soldered on a circuit board 1 cm away from each other. A large glass plate was held parallel to the circuit board 7 and 40 cm away from the ultrasonic transducers as a reflector of the ultrasonic wave. The resulting amplitudes of the acquired signals were 2.3 and 0.424 Vpp, respectively. These two simple tests verify the capability of the selected transducer pair to obtain a sufficiently large signal compared with the environmental noise level during the subsequent hand gesture experiment performed approximately 10 to 30 cm away from the transducer pair.

3.2. Circuit for Air Sonar Operation

The air sonar was designed with a transmitted frequency of 40 kHz. The minimum sampling rate should be greater than 80 kHz to acquire the reflected signal directly. It can be even higher than 200 kHz to obtain a better waveform of the reflected signal. The sampling rate and data amount for algorithm processing should be reduced to make the hand gesture recognition technology easily implementable online.

As we utilized the Doppler effect of the received signal for our hand gesture recognition, the sampling frequency is much lower than that required to acquire the signal emitted from the ultrasonic emitter. This could be realized through a frequency-mixing technique. Consider $f1$ as the frequency of the received ultrasonic wave, which is given by the frequency of the transmitted ultrasonic wave plus the Doppler frequency shift. Consider $f2$ as the frequency of the mixed signal, as specified. When signals of frequencies $f1$ and $f2$ enter the mixer, the mixed signals of frequencies $f1 + f2$ and $f1 - f2$ emerge at the output terminals. We employed the signal with the frequency ($f1 - f2$) for investigating the Doppler effect.

Based on the preliminary experiment, the magnitude of the Doppler frequency shift of interest is less than 500 Hz. We constructed the mixed signal with the frequency $f1 - f2$ = $f1$ (40.8 kHz + Doppler frequency shift) $- f2$(37.6 kHz) = 3.2 kHz. The frequency of 3.2 kHz was selected, as it was five times higher than the Doppler frequency shift in this study.

The oscillators, mixer, and filters were employed to construct the circuitry for reducing the carrier frequency [24]. Figure 4 shows the circuit implementation. Two oscillators were realized using a Wien bridge sine-wave oscillator. One oscillator was operated at 40.8 kHz ($f1$) to drive the ultrasonic emitter, and the other oscillator was operated at 37.6 kHz ($f2$), as the mixed signal. The analog multiplier AD633 served as the mixing function. A bandpass filter was applied to remove unwanted noises and obtain a reasonable frequency bandwidth signal. A low-pass filter was employed to remove the signal with a carrier frequency of $f1 + f2$. In addition, a voltage follower was added to avoid the loading effect of the voltage signal from the ultrasonic receiver to the bandpass filter. After bandpass filtering, a voltage amplifier was used to adjust the gain so that the signal sent to the mixer had a proper amplitude during the hand gesture operation.

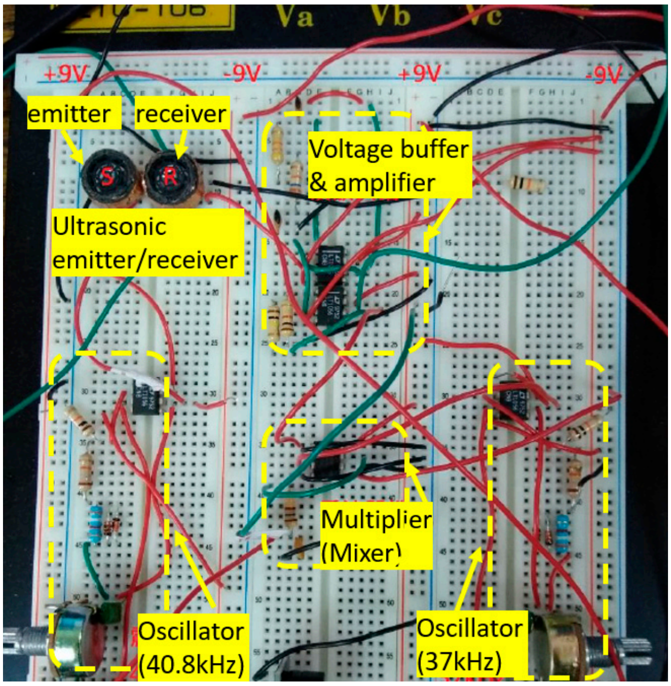

Figure 4. The circuit implementation for air sonar operation.

4. Detecting the Occurrence of the Hand Gesture Event

4.1. Ultrasonic Emitter and Receiver

To detect the occurrence of hand gesture event, we utilized the continuous short-period analysis. Each short-period was set as 0.02 s, which corresponded to 1000 data points based on a sampling rate of 50 kHz. The time resolution for detecting the occurrence of a hand gesture would be one-fiftieth of a second. The spectrum was analyzed by fast Fourier transform (FFT).

Two signal processing techniques were applied in this study: bandpass and notch filtering. The bandpass filter was employed to process the signal to remove the frequency bands other than that from 2.7 to 3.7 kHz. We chose a Butterworth filter of order 7 with the cutoff frequencies of 2.7 and 3.7 kHz to design the bandpass filter. The filtered time signals can be easily used to distinguish between no hand gesture and the occurrence of a hand gesture through an amplitude change.

Moreover, considering the critical moment at which the hand gesture started, we observed that the frequency peaks split into two major frequencies, with one at a carrier frequency of ~3200 Hz and the other deviating from the carrier frequency. The strength of the latter peak had a varied amplitude, which was related to the hand gesture executed. Once the algorithm for detecting the hand gesture starting time was to select the characteristic frequency as the peak frequency with the maximum amplitude for FFT results of each short-period (1000 sample points), the carrier frequency was possible to be selected as the characteristic frequency. Practically, the carrier frequency should have a narrower bandwidth compared to that of the Doppler frequency because the carrier frequency was synthesized by the constructed circuit. The Doppler frequency could easily cover a wider frequency range because performing the hand gesture was hard to maintain a constant velocity at all points of hand to reflect the ultrasonic waves. Thus, a proper notch filter design to effectively reduce the amplitude of carrier frequency peak and still allow the Doppler frequency peak to be selected as the characteristic frequency during hand gesture operation was needed.

By closely examining the amplitude of the main lobe for the case of no hand gesture, we considered a value of 0.007 as a reference for the notch filter design. The notch filter was set a center frequency as the carrier frequency. The bandwidth and the maximum attenuation gain of the notch filter were 5 Hz and 0.01, respectively.

Figure 5 shows the processed results of the time-domain signals and their frequency spectrum without any hand gesture and with a hand gesture (push motion) occurring approximately at 5 s. The spectrum results were overlaid by fast Fourier transform analysis through a continuous short-period analysis.

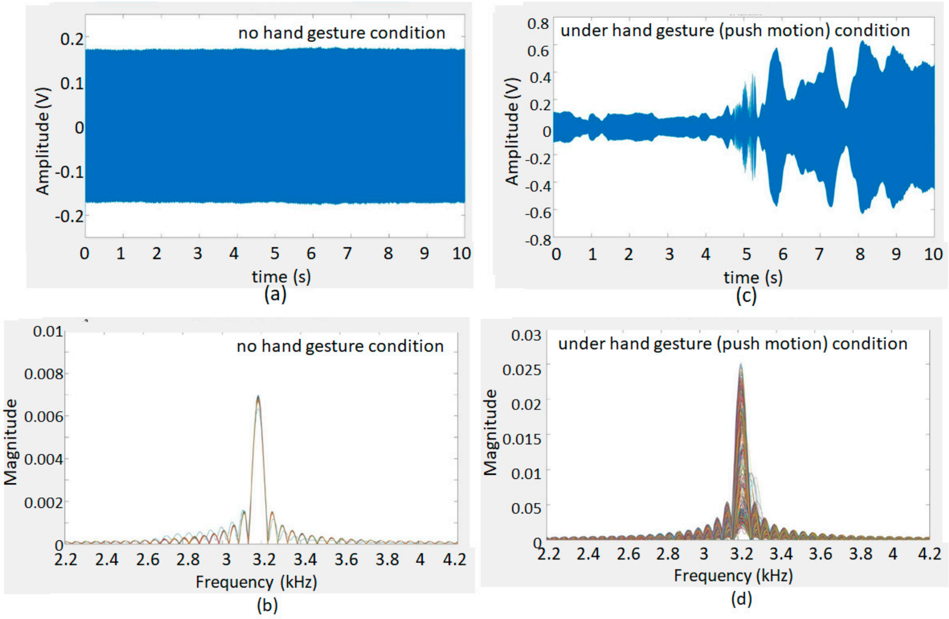

Figure 5. The acquired signal after bandpass and notch filtering: (**a**) time-domain signal and (**b**) its continuous FFT results of each 20 ms short-time period at no hand gesture condition. Time-domain signal (**c**) and its continuous FFT results of each 20 ms short-time period (**d**) under hand gesture (push motion) condition. The spectrum results were overlaid by fast Fourier transform analysis through a continuous short-period analysis.

4.2. Scheme of Judging the Start Time of Hand Gesture

Two parameters were investigated to determine the start time of the hand gesture operation for each short-time signal processed with the aforementioned bandwidth and notch filters: the frequency shifts and the amplitudes of the frequency peaks.

Figure 6a shows the frequency values of the frequency peaks obtained with each time frame of 0.02 s. These frequency values were first at a near constant level, and then showed a rapid increase and subsequent drop from a maximum. The time interval of the evident upsurge and decrease region was approximately 1 s, which indicated that the hand gesture started from an acceleration state, then changed to a deceleration one, and then stopped.

Although checking the maximum peak frequencies could determine the start time for most of the studied hand gestures, a likely situation is that the maximum peaks occurred around the designed cut-off frequencies of the bandpass filter without hand gesture motion. This was because the amplitude of the carrier frequency after processing with the notch filter could be less than the amplitudes of the peaks around the frequency region close to the cut-off frequency of the Butterworth filter. This would produce the peak-frequency curve jump from the carrier frequency to the frequency near the cut-off frequency of the bandpass filter and thus cause a mistake in judging the start time of the hand gesture.

In contrast, a more reliable method could be the use of the amplitudes of the maximum peaks to judge the start time of the hand gestures. Figure 6b shows the amplitudes of the maximum peak obtained with each time frame of 0.02 s. The amplitudes of the maximum peaks resulted in a larger variation during a hand gesture compared with the cases of a hand gesture at a standby position or no hand gesture. We first attempted to utilize the amplitude difference of the maximum peaks between neighboring time frames to judge the start time of the hand gestures. This method is considered as the amplitude differentiation of continuously maximum peaks. The result displayed more sharp peaks compared with the case without differentiation. For other hand gestures, this differentiation value was not sufficiently large to be selected as the representative start time of hand gestures, such as hand rotation motion. Consequently, we then performed the second differentiation of the result. The results of the second derivative were even sharper than those of the first derivative for all the hand gestures (Figure 6c).

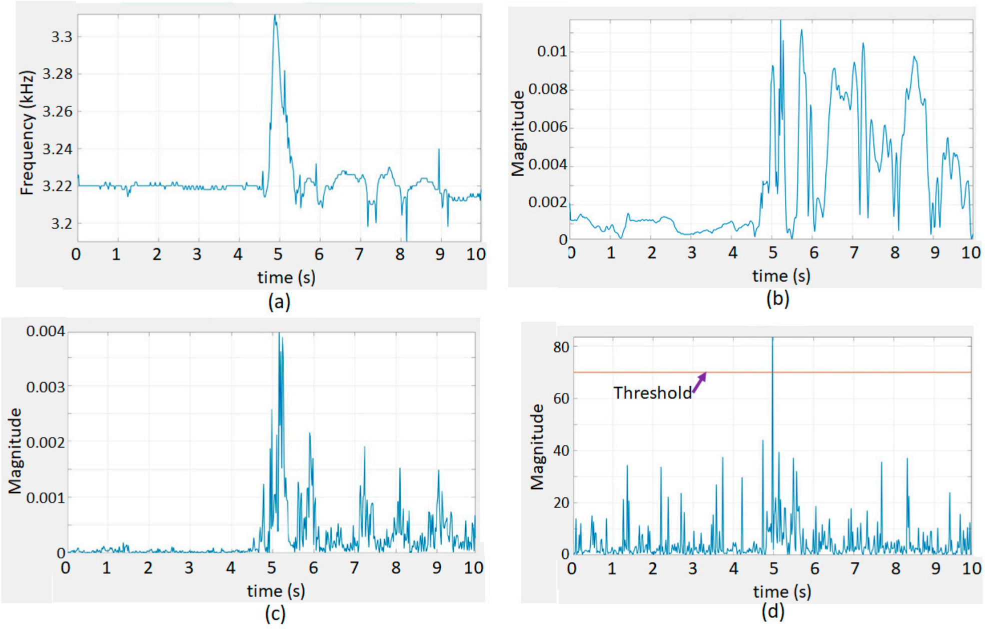

Figure 6. (**a**) The frequencies of the maximum peaks under a push hand gesture motion. (**b**) The amplitudes of the maximum peaks analyzed with the signal as (**a**). (**c**) The signal in (**b**) after 2nd differentiation. (**d**) The result of T_{pick} $[n]$ used to judge the start time of hand gesture.

Finally, we selected the start time T_{pick} $[n]$ of the hand gesture based on the second derivative of the strength at the peak frequency with the maximum strength by defining the following parameter:

$$T_{pick}[n] = \left| \frac{A[n+2]}{A[n+1] - A[n]} \right|, \tag{3}$$

where $A[n]$ is the second derivative of the amplitudes of the maximum peaks at time n.

If we consider that the amplitude of the maximum peak as a function of the hand position, $A[n]$ indicates the acceleration of the hand and the difference $A[n+1] - A[n]$ indicates the jerk of the hand. Therefore, in terms of physical meaning, we determine the moment of occurrence of minimum jerk and a high acceleration as the start time of the hand gesture. A threshold value could be determined to find the appropriate T_{pick} $[n]$ so that the time step n could be evaluated as the start time of the hand gesture (Figure 6d).

5. Scheme of Hand Gesture Recognition

5.1. Convert Motion Signal to Time-Frequency Response

After judging the start time of the hand gesture, we need to evaluate the acquired signal further for hand gesture recognition. We proposed a method of utilizing the time–frequency response to evaluate gestures. Figure 7 shows the block diagram of the proposed method. As the common gesture motions were performed for approximately 1 s, we could pick only the signal data in this time interval for performing the analysis of the time–frequency response once the start of the hand gesture was judged. This effectively reduced the electrical power of the microprocessor required for implementing this technique on a portable device. In this investigation, we took 150,000 sample points, which corresponded to a signal of a duration of 3 s, for spectrogram analysis while monitoring the occurrence of hand gestures. More specifically, once we detected the data point (=i) representing the start time of the hand gesture, the data points ranging from (i − 69,999) to (i + 80,000) were extracted for the spectrogram analysis. Considering the data length of 150,000 was a conservative choice to analyze a hand gesture motion fully because the resulting spectrogram, which is described in the following section, indicated that an effective hand gesture pattern only occupied approximately one half of the spectrogram in the x direction (time axis). The data length for judging one gesture could be significantly reduced by employing this technique in future real-time applications.

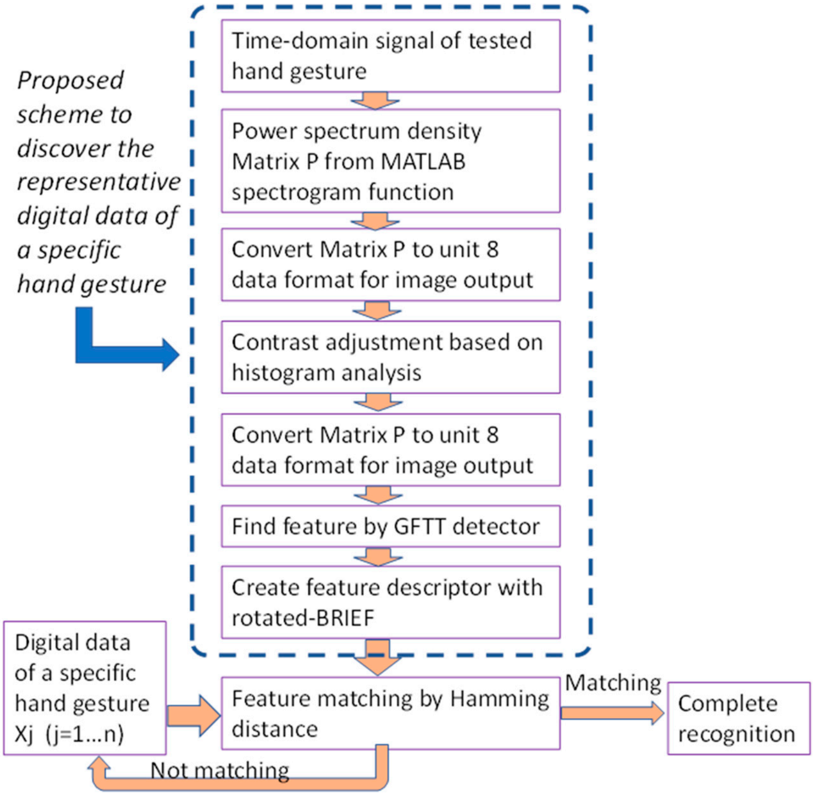
Proposed scheme to discover the representative digital data of a specific hand gesture
Time-domain signal of tested hand gesture
Power spectrum density Matrix P from MATLAB spectrogram function
Convert Matrix P to unit 8 data format for image output
Contrast adjustment based on histogram analysis
Convert Matrix P to unit 8 data format for image output
Find feature by GFTT detector
Create feature descriptor with rotated-BRIEF
Digital data of a specific hand gesture Xj (j=1…n)
Feature matching by Hamming distance
Matching
Complete recognition
Not matching

Figure 7. The block diagram of the proposed method for hand gesture recognition.

The analysis of the time–frequency response utilized the short-time Fourier transform (STFT). The two STFT parameters important for our analysis are the window and overlap sizes. The window size affected the resolution in the time-domain variation. Although a small window size yielded a better time resolution, the frequency-domain resolution was poor. The overlap size was also critical for the resolution in both the time and frequency domains. A smaller overlapping resulted in a worse resolution, which indicated that the pixelated image appeared in the spectrogram. However, a smoother spectrogram with a larger overlapping required a considerable calculation effort. A Hamming window with a window size of 20,000 and an overlap size of 19,000 was selected in this study. This indicates that the spectrogram derived for our hand gesture recognition had a time resolution of 0.02 s and a frequency resolution of 2.5 Hz.

5.2. Processing Spectrogram Result of the Image for Hand Gesture Recognition

As previously mentioned, STFT was applied to obtain the spectrogram for the signal of interest. The Hamming window size was set as 20,000, and the overlap length was set as 19,000 for analyzing the 150,000 data points. The image obtained from the surf function with the spectrogram computed results using the MATLAB software was directly used to obtain the plot with the frequency in the range of 2.7 to 3.7 kHz, as shown in Figure 8a. This image has two major characteristics: (1) the calculated power spectrum is in a double-precision float-point format; (2) the edge of the signal pattern appears insufficiently smooth. The former characteristic could make our proposed method of image pattern recognition more complicated and time-consuming. The latter characteristic could be an issue while finding the feature points.

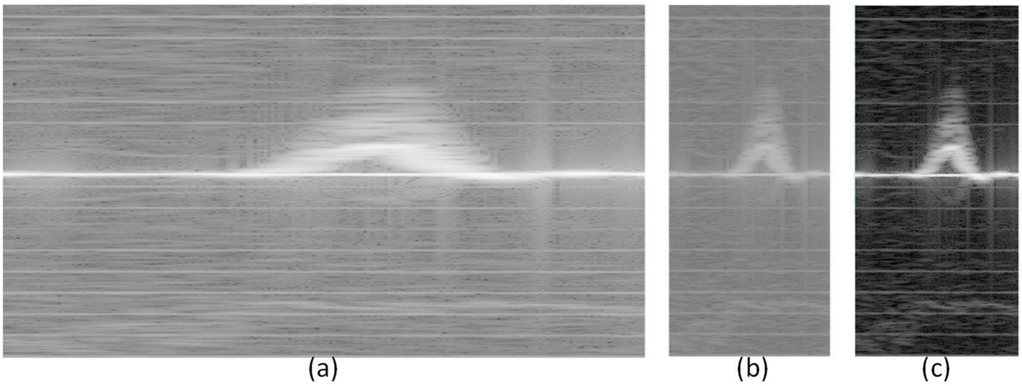

Figure 8. (**a**) The image of push-motion spectrogram obtained by the MATLAB surf function. (**b**) The image (**a**) was converted into uint8 format. (**c**) The image (**b**) processed with contrast adjustment.

Instead of using the default image output with surf function for the STFT result, we processed the output data of the image as follows. We converted the calculated power spectrum from the STFT format into the uint8 format, that is, all the values of the power spectrum were shown as integers from 0 to 255, and then displayed the resulting values as a grayscale image. During this conversion, only the frequency range of the STFT data was processed, which ranges from 2.7 to 3.3 kHz. This reduced the processing time of the calculation compared with dealing with the entire frequency range of the STFT results from 0 to 25 kHz. Figure 8b shows the resulting image of the push motion.

However, the contrast of the image was not sufficient to differentiate the motion part from the background. An image processing technique for histogram equalization was employed to enhance the contrast. Based on the histogram analysis of the image derived from the STFT, all the pixels were at a higher gray level, which indicates that the image had an over-exposure trait. We utilized the following equations so that the pixel values could be mapped in the range from 0 to 255:

$$\mathrm{s} = \begin{cases} 0, \text{ if } r < r_1 \\ \frac{255r}{255-r_1} - \frac{255r_1}{255-r_1} \end{cases}, \tag{4}$$

where r is the gray level of the original image and s is the gray level of the adjusted image. r_1 is the threshold value used in this study, which is 170.

Figure 8c shows the contrast-adjusted image. Its histogram shows that the pixel values occupy the full range of gray level. The processed contrast-adjusted images of varied gestures, which are referred to as featured images hereinafter, will be investigated below.

5.3. Image Recognition of Featured Gesture Spectrogram

Four gestures were investigated to demonstrate gesture identification with our proposed featured image method. We used the detector offered in OpenCV (Open Source Computer Vision Library) to find the corners of the featured images (keypoint estimation) and then computed a descriptor for the detected keypoints. Subsequently, the feature match function provided by MATLAB was employed to determine the number of matched pairs. A threshold value could be determined to judge a test gesture for the best fitting of candidate gestures.

5.3.1. Monitor Features Using the GFTT Detector

The corners are critical characteristics of an image. The corners could be considered keypoints with the maximal values of the first-order derivative along all the directions. We first employed the Harris corner detection method [25]. For an image represented by the pixels in the x and y directions, an autocorrelation function $E(u, v)$ that determines the intensity variation at the center point (x, y) with its neighborhood window Q is given by

$$E(u,\ v) = \sum\nolimits_{x,y} w(x,y)[I(x+u,\ y+v) - I(x,y)]^2, \tag{5}$$

where $w(x, y)$ is a window function, (u, v) represents the shifts in the x and y directions, respectively, and $I(x + u, y + v)$ is the intensity at the position $(x + u, y + v)$. $E(u, v)$ allows us to find the window Q with a large variation in both the x and y directions. However, this calculation is relatively time-consuming. Alternatively, Taylor's expansion could be utilized to obtain the approximated result:

$$I(x+u,\ y+v) \approx I(x,y) + \frac{\partial I(x,y)}{\partial x}u + \frac{\partial I(x,y)}{\partial y}v.$$

Therefore, we could simplify $E(u, v)$ as follows:

$$E(u,\ v) = \sum\nolimits_{x,y} w(x,y)\left(I_x u + I_y v\right)^2. \tag{6}$$

Furthermore, we can rewrite the above equation as

$$E(u,\ v) = [u\ v](\sum w(x,\ y)\begin{bmatrix} I_x^2 & I_x I_y \\ I_x I_y & I_y^2 \end{bmatrix})\begin{bmatrix} u \\ v \end{bmatrix}. \tag{7}$$

The matrix characterizes the structure of the gray levels at a given point. Considering the eigenvalues of the matrix as λ_1 and λ_2, we could categorize $E(u, v)$ into three groups as follows: (1) both λ_1 and λ_2 are small, indicating that the intensity of the neighborhood window of interest is nearly constant; (2) if one of the eigenvalues is greater than the other, indicating an edge existence; (3) if both λ_1 and λ_2 are large, indicating a corner occurrence.

The GFTT detector evaluates a corner by utilizing the cornerness scoring function, which is defined as

$$R = \min(\lambda_1,\ \lambda_2). \tag{8}$$

If the minimum eigenvalue is larger than a predefined threshold, then the neighborhood window is considered a corner. In this study, we used the function goodFeaturesToTrack with its default parameters given in the OpenCV database to find the feature corners of the STFT results.

5.3.2. Feature Descriptor with Rotated BRIEF and Feature Matching

We employed rotated BRIEF as the feature descriptor. Rotated BRIEF is based on BRIEF, which only compares the intensity between two pixel positions around the detected feature points to build a binary descriptor. Moreover, matching the binary descriptors only requires the computation of the Hamming distances through XOR with a very fast speed [26]. To mitigate the limitation of BRIEF, steered BRIEF, which is helpful in increasing the orientation invariance, can be used. However, steered BRIEF exhibits a limitation in differentiating the feature descriptors of different feature points.

We then selected rotated BRIEF for a better discerning ability of the feature descriptor. Rotated BRIEF attempted to find the optimal sampling pairs, in contrast to BRIEF, which uses randomly chosen sampling pairs, for the binary intensity tests. The detailed algorithm of rotated BRIEF is as follows [18,27]:

First, the test set for N ($\geq$256) feature points was created. For each feature point, 31 × 31 neighboring pixel points were taken as a large patch P1 to perform Gaussian smoothing. Patch W2 of arbitrary 5 × 5 points was selected among W1, and the gray levels of these 25 points were averaged as the value of a representative point of W2. Thus, the total number of points, i.e., (31 − 5 + 1) × (31 − 5 + 1) = 729 of W2, formed the representative points. To extract these representative points as point pairs, we need M ($C_2^{729} = 729 \times 728 \div 2 = 265,356$) arrangements.

Subsequently, the point pairs at positions x and y were compared based on M arrangements. The comparison followed the BRIEF scheme with

$$\tau(p;\, x,\, y) = \begin{cases} 1 & \text{if } p(x) < p(y), \\ 0 & \quad \text{otherwise} \end{cases} \tag{9}$$

where $p(x)$ is the intensity value at position x.

A matrix Q of size N × M was created. Each row of the matrix Q is a binary vector for the feature descriptor. We obtained the average of each row of Q and rearranged the row vector according to the distance between the average of each row of Q and 0.5. The re-sequenced row vectors formed the matrix T.

This was followed by a greedy search. The first-row vector of T was placed into the first row of matrix R. The next-row vector from T was compared with all the row vectors in R. If its absolute correlation was greater than a threshold, the vector was discarded; else, it was added to R. The previous step was repeated until there were 256 row vectors in R. If the resulting row vectors of R were fewer than 256, the threshold was increased and the above process was repeated.

Finally, we matched two sets of binary feature vectors using the Hamming distance. The MATLAB function matchFeatures was applied with a default threshold of 10. The threshold represents the percentage of distance from a perfect match. Two feature vectors were considered to be matched when the distance between them was less than the specified threshold.

6. Recognition Results of the Same Gestures and Different Gestures

Using the aforementioned methods, we investigated the four types of gestures described in the previous section. For each studied gesture, we randomly took two samples from five hand gesture operations. The samples were performed by the same person. We first obtained the recognition results for the same gestures. For each of the four studied gestures, we took the samples at different times as motions "A" and "B" and compared them with each other to obtain the matching points of the extracted features. Table 1 lists the test results of the matching points. Figure 9a shows that the results of the push motion "A" tested against itself yielded 245 matching points and the push motion "B" tested against itself yielded 251 matching points. The small yellow circles indicate the matching points. Figure 9b shows that, by using the push motion "A" to test the push motion "B", we could obtain 189 matching feature points. When we switched the matching sequence, that is, using the push motion "B" to the test motion "A", we could obtain 179 matching points. The pinch out motion "A" tested against itself yielded 210 matching feature points, and the pinch out motion "B" tested against itself yielded 240 matching points. Using the pinch out motion "A" to test "B", we could obtain 165 matching points, and we could obtain 149 matching points by switching the test order.

A similar test was also performed for the hand rotation motion. The rotation "A" tested against itself yielded 320 matching points, and rotation "B" tested against itself yielded 311 matching points. The cross-validation of using rotation "A" to test "B" yielded 192 matching points and the reverse order yielded 188 matching points. As for the gesture of wrist motion from flexion to extension, the motion "A" tested against itself yielded 267 matching points; the motion "B" tested against itself yielded 250 matching points. Using motion "A" to test "B" yielded 156 matching points, and the reverse order yielded 140 matching points.

Table 1. Results of matching points for the same and different hand gestures.

	Push A	Push B	Wrist A	Wrist B	Pinch A	Pinch B	Rot. A	Rot. B
Push A	245	189	18	16	27	24	26	32
Push B	179	262	12	11	26	23	48	36
Wrist A	19	17	267	156	37	28	25	26
Wrist B	17	12	140	249	19	23	19	31
Pinch A	26	23	26	20	210	165	33	43
Pinch B	24	20	11	19	149	232	34	35
Rot. A	30	28	31	28	37	45	320	192
Rot. B	44	32	21	30	38	33	188	288

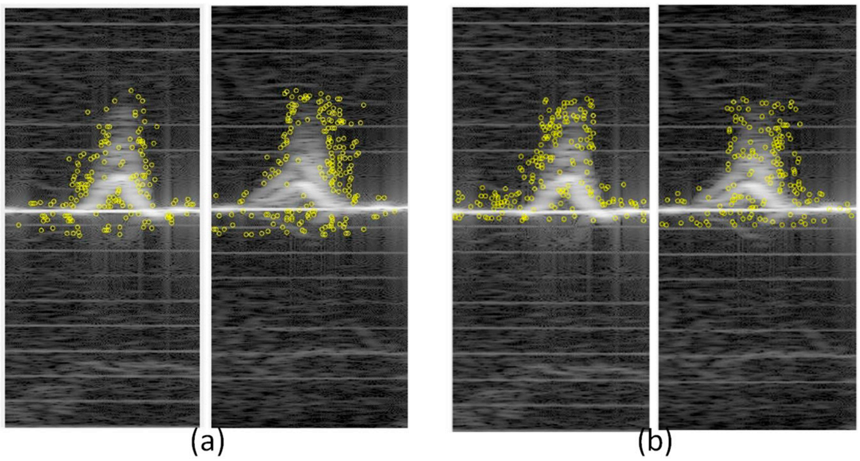

Figure 9. (**a**) (**Left**) The results of the push motion 'A' testing itself with 245 matching points (marked with yellow circles), and (**Right**) the push motion 'B' testing itself with 251 matching points. (**b**) The push motion 'A' to test the push motion 'B' with 189 matched feature points.

From the above analysis, we observe that a stronger matching occurs at the comparison of each identical feature description for each investigated case. The rotation motion case showed the highest matching number of 320 points, and the pinch out motion case showed the lowest matching number of 210 points. Although the matching points of cross-validation for the same gestures were not as high as those for testing with an identical gesture, they still exhibited the largest value of 192 and the smallest value of 140 for the four investigated gestures. In addition, we observed that the comparison order could affect the matching number of points for the same gesture motions. The differences of cross-validation comparisons were 10, 16, 16, and 4 for push motion, wrist motion from flexion to extension, pitch out, and rotation, respectively. The disparity is approximately one order of magnitude less than the number of matching points, which indicates that the effect of the test order could be reasonably neglected.

Subsequently, different gestures were examined. For every two cases of the same gestures discussed above, we tested the other three cases of gestures. We selected some examples to illustrate the matching results. Figure 10a–e shows the results of using the rotation motion "A" to test the push motion "A", the wrist motion from flexion to extension "A" to test the pinch out motion "A", the wrist motion from flexion to extension to test the rotation motion "A", the pinch out motion "A" to test the wrist motion from flexion to extension "A", and the push motion "A" to test the rotation motion "A", respectively. The corresponding numbers of matching points were 30, 37, 25, 26, and 26, respectively. The marked matching points (yellow circle) in Figure 10 indicate much smaller numbers compared with that of the same gestures. Therefore, a borderline of the matching points could be set to distinguish similar gestures and different gestures using our proposed method.

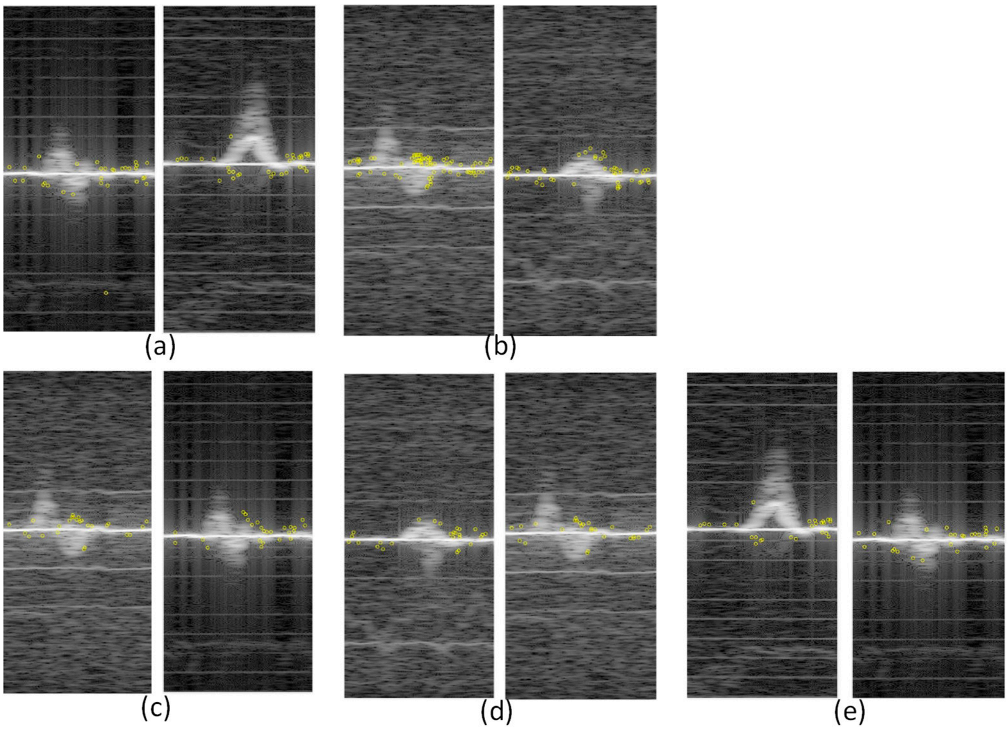

Figure 10. The results of feature points matching: (**a**) using the rotation motion 'A' to test push motion 'A'. (**b**) The motion of wrist from flexion to extension 'A' to test pinch out motion 'A'. (**c**) The motion of wrist from flexion to extension to test the rotation motion 'A'. (**d**) The pinch out motion 'A' to test the motion of 'wrist from flexion to extension. (**e**) The push motion 'A' to test the rotation motion 'A'.

To correlate the hand gestures and matching results quantitatively, we categorized the matching results into three groups. The first group consisted of identical hand gestures with the number of matching points between 320 and 210. The second group had the same hand gestures, but they were performed at different times. The number of matching points was between 140 and 189. The third group consisted of different hand gestures, and the number of matching points was between 48 and 11.

If we define the maximum number of matching points in the aforementioned cases, which is 320, as a denominator to find the matching ratio of hand gesture recognition, then the identical hand gestures are indicated by a matching ratio beyond 0.66 (i.e., greater than 210 divided by 320). The same gestures are indicated by matching ratios above 0.44. The different hand gestures could be determined by a matching ratio of less than 0.15. The results indicate that there is a large ratio interval of 0.29 to prevent the misjudgment of hand gestures.

Furthermore, if we directly find the probability density functions according to the experimental matching points shown in Table 1 for the same hand gestures and the different hand gestures, the probability of accuracy rate could be found. The mean value (μ) and standard deviation (σ) are 214.44 and 53.43 for the identical gestures, 26.73 and 9.02 for the different gestures. Using the probability of normal distribution, the accuracy rate could be estimated as below:

Let X_1 be the matching points for the cases of failure of hand gesture recognition possessing the normal distribution $N(\mu_1, \sigma_1^2)$ with probability density function (pdf) $F_1(x)$ and cumulative density function (cdf) $F_1(x)$ and X_2 be the matching points for the cases of correct hand gesture recognition possessing the normal distribution $N(\mu_2, \sigma_2^2)$ with pdf

$F_2(x)$ and cdf $F_2(x)$, $\mu_1 < \mu_2$. The area of intersection zone, which indicates the probability of faulty hand gesture recognition, could be found by

$$\begin{aligned} &P(X_1 > c) + P(X_2 < c) \\ &\quad = 1 - F_1(c) + F_2(c) \\ &\quad = 1 - \tfrac{1}{2}\mathrm{erf}\left(\tfrac{c-\mu_1}{\sqrt{2}\sigma_1}\right) + \tfrac{1}{2}\mathrm{erf}\left(\tfrac{c-\mu_2}{\sqrt{2}\sigma_2}\right) \end{aligned}, \tag{10}$$

where erf(.) means the error function, c is the x-value for $F_1(x) = F_2(x)$ and can be obtained by

$$c = \frac{-\sigma_2^2\mu_1 - \sigma_1^2\mu_2 + \sigma_1\sigma_2\sqrt{\left(2(\sigma_2^2 - \sigma_1^2)\log\left(\frac{\sigma_2}{\sigma_1}\right) + (\mu_2 - \mu_1)^2\right)}}{(\sigma_1^2 - \sigma_2^2)} \tag{11}$$

Figure 11 shows the analyzed results. The accuracy rate of hand gesture recognition could achieve a probability of 99.8% by using the developed scheme.

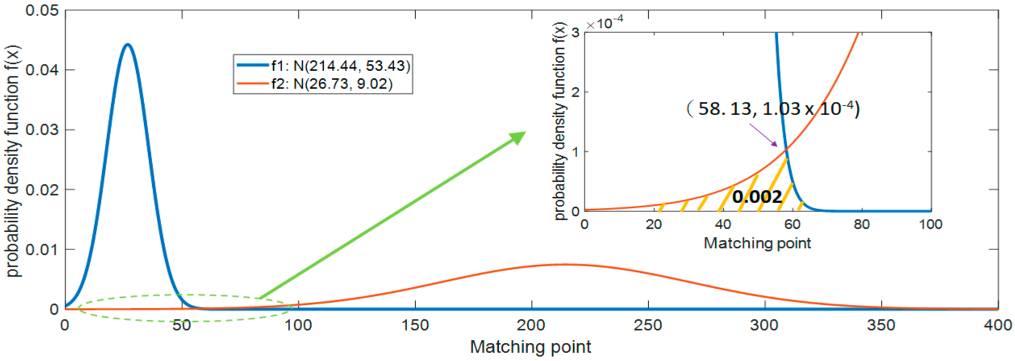

Figure 11. The probability density functions based on experimental results to find the accuracy rate by using the proposed scheme of hand gesture recognition.

We estimated the required time to execute the proposed algorithm for hand gesture recognition. The calculation was based on the MATLAB code operated in the personal computer with the hardware configuration of Intel Core i5 CPU @3.10 GHz and 8 GB RAM. Through the MATLAB timer functions of tic (starts a stopwatch timer) and toc (prints the elapsed time since tic was used). The amount of time required to complete one hand gesture recognition was about 0.7 s. The computational cost could be further reduced by converting the MATLAB code to C program language or Python language for better performance.

In the future, we will explore more different types of hand gestures for recognition. More data from different testers will be analyzed. In addition, the parameters used in the algorithm could be further investigated to obtain optimized results. For example, the window functions used in the STFT analysis and the threshold values in the MATLAB matchFeatures function.

The proposed methods in this study could be applied to other problems such as structural health monitoring or fault diagnosis of machines [28–30]. For example, using the receiving acoustic signal along with the presented signal processing scheme possesses the advantages of low-cost hardware setup and non-destructive detection. The described image processing scheme could be also employed for thermal imaging data analysis. For instance, using a specific fusion method to extracting features, along with nearest neighbor classifier and support vector machine has been studied as an effective way for fault diagnosis of the angle grinder [30]. It could be interesting for further investigation by using our proposed imaging processing scheme.

7. Conclusions

Based on linear and rotational Doppler effect, four different hand gestures of push, wrist motion from flexion to extension, pinch out, and hand rotation were studied for gesture recognition. The hardware of the ultrasonic monitoring system was implemented with a circuit containing a frequency mixer, filters, and oscillators to lower our sampling frequency and acquire Doppler signals. An algorithm to judge the start time of the hand gestures was proposed by obtaining the second derivative of the strength at the peak frequency with the maximum strength. Hand gesture recognition was performed using the image constructed using the STFT results. The GFTT algorithm was employed to find the feature corners of the images of the STFT results. A rotated BRIEF feature descriptor was used to perform the binary coding of the feature corners, and the Hamming distance was employed to match the feature descriptor. Based on the experimental results, the accuracy rate of hand gesture recognition with the proposed scheme reaches 99.8%.

Author Contributions: Conceptualization, G.-H.F.; methodology, G.-H.F. and G.-R.L.; software, G.-R.L.; data curation, G.-R.L.; writing—original draft preparation, G.-H.F. and G.-R.L.; writing—review and editing, G.-H.F.; project administration, G.-H.F.; funding acquisition, G.-H.F. Both authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Ministry of Science and Technology in Taiwan, grant no. 106-2221-E-194-035-MY3. The APC was funded by Ministry of Science and Technology in Taiwan.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Acknowledgments: The authors would like to thank the Ministry of Science and Technology in Taiwan for financial supports under contract no. 106-2221-E-194-035-MY3.

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

References

1. Geng, J.; Xie, J. Review of 3-D endoscopic surface imaging techniques. *IEEE Sens. J.* **2013**, *14*, 945–960. [CrossRef]
2. Xu, Y.; Wang, Y.; Yuan, J.; Cheng, Q.; Wang, X.; Carson, P.L. Medical breast ultrasound image segmentation by machine learning. *Ultrasonics* **2019**, *91*, 1–9. [CrossRef]
3. Biswas, A.; Abedin, S.; Kabir, M.A. Moving Object Detection Using Ultrasonic Radar with Proper Distance, Direction, and Object Shape Analysis. *J. Inf. Syst. Eng. Bus. Intell.* **2020**, *6*, 99–111. [CrossRef]
4. Feng, G.H.; Liu, H.J. Piezoelectric Micromachined Ultrasonic Transducers with a Cost-Effective Bottom-Up Fabrication Scheme for Millimeter-Scale Range Finding. *Sensors* **2019**, *19*, 4696. [CrossRef] [PubMed]
5. Yan, J.; Yang, X.; Sun, X.; Chen, Z.; Liu, H. A Lightweight Ultrasound Probe for Wearable Human–Machine Interfaces. *IEEE Sens. J.* **2019**, *19*, 5895–5903. [CrossRef]
6. Liu, H.; Wang, L. Gesture recognition for human-robot collaboration: A review. *Int. J. Ind. Ergon.* **2018**, *68*, 355–367. [CrossRef]
7. Oudah, M.; Al-Naji, A.; Chahl, J. Hand gesture recognition based on computer vision: A review of techniques. *J. Imaging* **2020**, *6*, 73. [CrossRef]
8. Feng, G.H.; Liu, H.J.; Lai, G.R. Piezoelectric Micromachined Ultrasonic Transducer with a Universal Bottom-Up Fabrication Approach Implemented on a Foil as Doppler Radar for Gesture Recognition. In Proceedings of the 2019 IEEE 32nd International Conference on Micro Electro Mechanical Systems (MEMS), Seoul, Korea, 27–31 January 2019; pp. 779–782.
9. Khan, F.; Leem, S.K.; Cho, C.H. Hand-based gesture recognition for vehicular applications using IR-UWB radar. *Sensors* **2017**, *17*, 833. [CrossRef]
10. Liu, X.; Li, K.; Liu, D.C. A sound-based gesture recognition technology designed for mobile platform. *J. Inf. Comput. Sci.* **2015**, *12*, 985–991. [CrossRef]
11. Gupta, S.; Morris, D.; Patel, S.; Tan, D. Soundwave: Using the doppler effect to sense gestures. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Austin, TX, USA, 5–10 May 2012; pp. 1911–1914.
12. Przybyla, R.J.; Tang, H.Y.; Guedes, A.; Shelton, S.E.; Horsley, D.A.; Boser, B.E. 3D ultrasonic rangefinder on a chip. *IEEE J. Solid-State Circuits* **2015**, *50*, 320–334. [CrossRef]
13. Zhou, F.; Li, X.; Wang, Z. Efficient High Cross-User Recognition Rate Ultrasonic Hand Gesture Recognition System. *IEEE Sens. J.* **2020**, *20*, 13501–13510. [CrossRef]
14. ping Tian, D. A review on image feature extraction and representation techniques. *Int. J. Multimed. Ubiquitous Eng.* **2013**, *8*, 385–396.
15. Lowe, D.G. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* **2004**, *60*, 91–110. [CrossRef]

16. Azhar, R.; Tuwohingide, D.; Kamudi, D.; Suciati, N. Batik image classification using SIFT feature extraction, bag of features and support vector machine. *Procedia Comput. Sci.* **2015**, *72*, 24–30. [CrossRef]
17. Bay, H.; Ess, A.; Tuytelaars, T.; van Gool, L. Speeded-up robust features (SURF). *Comput. Vis. Image Underst.* **2008**, *110*, 346–359. [CrossRef]
18. Rublee, E.; Rabaud, V.; Konolige, K.; Bradski, G. ORB: An efficient alternative to SIFT or SURF. In Proceedings of the 2011 International Conference on Computer Vision, Barcelona, Spain, 6–13 November 2011; pp. 2564–2571.
19. Bhat, A. Makeup invariant face recognition using features from accelerated segment test and eigen vectors. *Int. J. Image Graph.* **2017**, *17*, 1750005. [CrossRef]
20. Mukherjee, D.; Wu, Q.J.; Wang, G. A comparative experimental study of image feature detectors and descriptors. *Mach. Vis. Appl.* **2015**, *26*, 443–466. [CrossRef]
21. Gupta, S.; Kumar, M.; Garg, A. Improved object recognition results using SIFT and ORB feature detector. *Multimed. Tools Appl.* **2019**, *78*, 34157–34171. [CrossRef]
22. Dutta, K.; Bhattacharjee, D.; Nasipuri, M.; Krejcar, O. Complement component face space for 3D face recognition from range images. *Appl. Intell.* **2021**, *51*, 2500–2517. [CrossRef]
23. Chen, V.C.; Li, F.; Ho, S.S.; Wechsler, H. Analysis of micro-Doppler signatures. *IEE Proc. Radar Sonar Navig.* **2003**, *150*, 271–276. [CrossRef]
24. Ultrasound Listener. Electronics That Expand Your Auditory Perception. Available online: https://www.instructables.com/Ultrasound-Listener-Electronics-that-expand-your-a/ (accessed on 16 May 2021).
25. Mouats, T.; Aouf, N.; Nam, D.; Vidas, S. Performance evaluation of feature detectors and descriptors beyond the visible. *J. Intell. Robot. Syst.* **2018**, *92*, 33–63. [CrossRef]
26. Calonder, M.; Lepetit, V.; Ozuysal, M.; Trzcinski, T.; Strecha, C.; Fua, P. BRIEF: Computing a local binary descriptor very fast. *IEEE Trans. Pattern Anal. Mach. Intell.* **2011**, *34*, 1281–1298. [CrossRef]
27. Wang, Y.; Zhang, S.; Yang, S.; He, W.; Bai, X. Mechanical assembly assistance using marker-less augmented reality system. *Assem. Autom.* **2018**, *38*, 77–87. [CrossRef]
28. Gelman, L.; Petrunin, I.; Parrish, C.; Walters, M. Novel health monitoring technology for in-service diagnostics of intake separation in aircraft engines. *Struct. Control Health Monit.* **2020**, *27*, e2479. [CrossRef]
29. Dubey, A.; Denis, V.; Serra, R. A Novel VBSHM Strategy to Identify Geometrical Damage Properties Using Only Frequency Changes and Damage Library. *Appl. Sci.* **2020**, *10*, 8717. [CrossRef]
30. Glowacz, A. Ventilation Diagnosis of Angle Grinder Using Thermal Imaging. *Sensors* **2021**, *21*, 2853. [CrossRef] [PubMed]