Journal of *Imaging*

MDPI

Review

A Structured and Methodological Review on Vision-Based Hand Gesture Recognition System

Fahmid Al Farid [1,*], Noramiza Hashim [1], Junaidi Abdullah [1], Md Roman Bhuiyan [1,*], Wan Noor Shahida Mohd Isa [1], Jia Uddin [2], Mohammad Ahsanul Haque [3] and Mohd Nizam Husen [4]

1 Faculty of Computing and Informatics, Multimedia University, Persiaran Multimedia, Cyberjaya 63100, Malaysia; noramiza.hashim@mmu.edu.my (N.H.); junaidi.abdullah@mmu.edu.my (J.A.); wan.noorshahida.isa@mmu.edu.my (W.N.S.M.I.)
2 Technology Studies Department, Endicott College, Woosong University, Daejeon 32820, Korea; jia.uddin@wsu.ac.kr
3 Department of Computer Science, Aarhus University, 9100 Aarhus, Denmark; iamahsanul@gmail.com
4 Cybersecurity & Technological Convergence, Malaysian Institute of Information Technology Universiti Kuala Lumpur, Kuala Lumpur 50250, Malaysia; mnizam@unikl.edu.my
* Correspondence: fahmid.farid@gmail.com (F.A.F.); romanbhuiyanpv@gmail.com (M.R.B.)

Abstract: Researchers have recently focused their attention on vision-based hand gesture recognition. However, due to several constraints, achieving an effective vision-driven hand gesture recognition system in real time has remained a challenge. This paper aims to uncover the limitations faced in image acquisition through the use of cameras, image segmentation and tracking, feature extraction, and gesture classification stages of vision-driven hand gesture recognition in various camera orientations. This paper looked at research on vision-based hand gesture recognition systems from 2012 to 2022. Its goal is to find areas that are getting better and those that need more work. We used specific keywords to find 108 articles in well-known online databases. In this article, we put together a collection of the most notable research works related to gesture recognition. We suggest different categories for gesture recognition-related research with subcategories to create a valuable resource in this domain. We summarize and analyze the methodologies in tabular form. After comparing similar types of methodologies in the gesture recognition field, we have drawn conclusions based on our findings. Our research also looked at how well the vision-based system recognized hand gestures in terms of recognition accuracy. There is a wide variation in identification accuracy, from 68% to 97%, with the average being 86.6 percent. The limitations considered comprise multiple text and interpretations of gestures and complex non-rigid hand characteristics. In comparison to current research, this paper is unique in that it discusses all types of gesture recognition techniques.

Keywords: gesture recognition; feature extraction; gesture classification; recognition accuracy; deep learning

Citation: Al Farid, F.; Hashim, N.; Abdullah, J.; Bhuiyan, M.R.; Shahida Mohd Isa, W.N.; Uddin, J.; Haque, M.A.; Husen, M.N. A Structured and Methodological Review on Vision-Based Hand Gesture Recognition System. *J. Imaging* **2022**, *8*, 153. https://doi.org/10.3390/jimaging8060153

Academic Editor: Mohamed Daoudi

Received: 15 April 2022
Accepted: 20 May 2022
Published: 26 May 2022

1. Introduction

1.1. Background

Hand gesture recognition plays a significant part in delivering diverse messages using hand gestures in the digital domain. Real-time hand gesture identification is now possible because of advancements in both imaging technology and image processing algorithmic frameworks. This has enabled natural interactivity previously unattainable by the use of the two-dimensional mouse. Due to the real-time nature of gesture recognition, it should be accomplished without overburdening the computing element. Moreover, image processing plays a critical role in segmentation, feature extraction of hand gestures and ultimate recognition of the gestures. Numerous computer vision algorithmic frameworks based on image processing concepts have been developed and are being improved.

Hand motions may vary from static to dynamic, depending on their use. Hand gesture recognition technologies each have their own set of benefits and drawbacks, which are dependent on the platforms on which they are implemented. Due to numerous difficulties encountered during foreground separation from the background, there are many current obstacles to achieving realistic and effective real-time hand gesture recognition. The hand that needs to be identified is represented by the foreground. Changing picture luminance, such as pixel color of the hand skin and background in vision-based systems, as well as cumbersome, expensive gear in glove-enabled and depth-enabled systems, are the most common problems.

1.2. Survey Methodology

The initial stage to the systemic literature review is to acquire all documents collected from 2012 to 2022, in which the screening procedure involves the download and lecture of the materials published in IEEE Xplore.

- Science, technology, or computer science are all acceptable search terms. Journals, proceedings, and transactions are the three main categories of publications.
- Article type is the in-depth analysis and commentary.
- The vision-based hand gesture recognition system can recognize a variety of different hand motions.
- The language of instruction is English.

The article selection process undertaken is illustrated in Figure 1. To identify relevant papers, we used three well-known databases: Scopus, Web of Science, and IEEE Xplore. We have also resorted to a number of reputable websites in order to obtain some particular information. The focus of this study is primarily determined by two keywords: hand gesture and computer vision. As a result, we utilized these two terms in conjunction with other keywords in the majority of our searches. Original articles, review articles, book chapters, conference papers, and lecture notes, among other kinds of materials, were gathered.

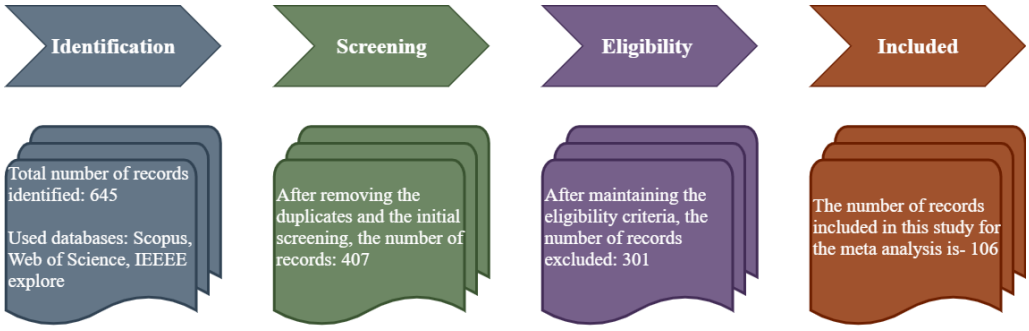

Figure 1. Block diagram of the article selection process.

We read the title, abstract, conclusion, and keywords in the first round of the review. Following the removal of duplicates, the surviving items are classified according to their intended use. During the first round of screening, hand gesture detection utilizing computer vision-based articles written in English are used as the selection criteria. The chosen papers are examined in the second stage of screening. The detailed approach serves as the ultimate criterion for article selection in this case. Only "108" articles are chosen for this evaluation after adhering to the aforementioned criteria. The number of chosen documents is divided into four categories: journal articles (62%), conference papers (28%), book chapters (6%), and others (4%), as shown in Figure 2. Figure 3 shows the number of publications of the chosen papers by year for this study. It may be argued that the greatest number of publications, 74, were chosen for this study during the past few years (after excluding irrelevant papers), indicating the significance of this area of research.

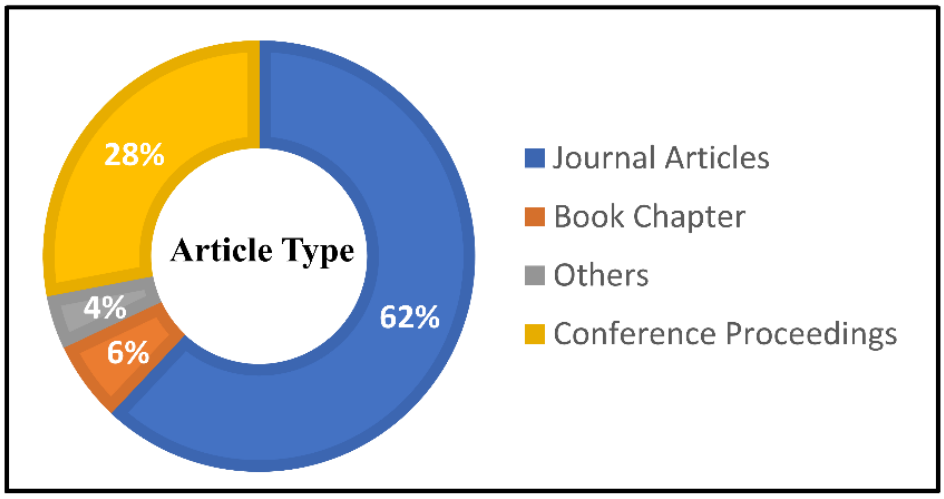

Figure 2. Journal articles, conference papers, book chapters, and other publications.

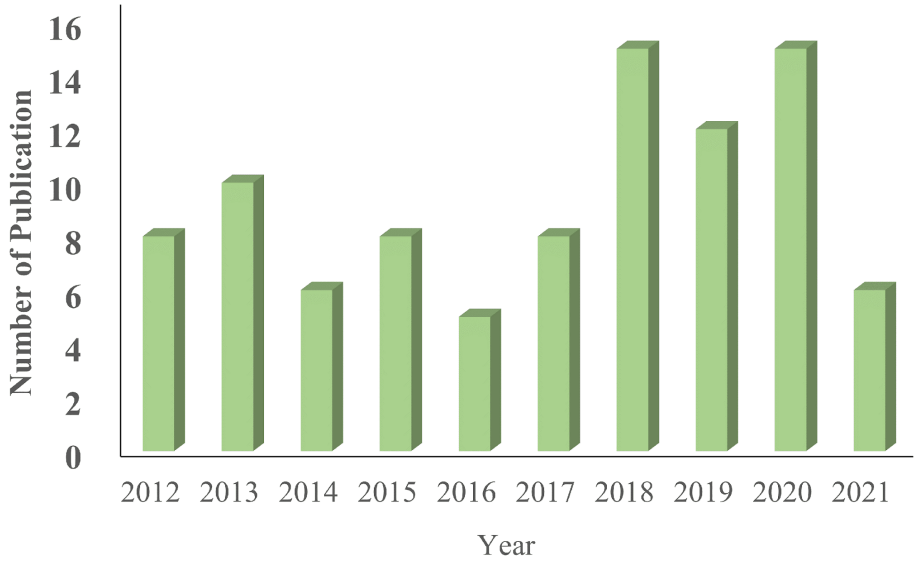

Figure 3. The number of peer-reviewed publications in relation to the year of publication.

1.3. Research Gaps and New Research Challenges

This section summarizes the recognized research gaps as well as the forthcoming research issues. Hand gesture recognition systems have prospective uses, according to the study. Furthermore, the region has been proven to have a number of obstacles from three perspectives: system, environment, and gesture-related issues. The vast diversity of possible gestures is the primary difficulty in vision-driven gesture detection. The identification of gestures requires a wide range of degrees of freedom, as well as a large amount of 2D variability. Dealing with different degrees of freedom, huge variety in the look of 2D, given the camera perspective (having the same gesture), varied silhouette sizes (such as spatial resolution), and varying resolutions in the time dimension are all part of the gesture identification process (such as gesture speed variability). Regardless of the application type, solution cost, or other variables, scalability, robustness, and user independence must be balanced for the best balance of accuracy, performance, and effectiveness. This should be possible if the system is capable of evaluating incoming video frames and responding to recognized gestures in real time. The requirement for resilience is one of the most essential elements for robustness in the identification of various hand motions in various light and busy situations. It is also crucial that the system can rotate pictures in-plane

and out-of-plane without breaking. Furthermore, scalability ensures that a large gesture vocabulary, consisting of just a few fundamental motions, can be handled. In this respect, the user has complete control over the composition of user gestures. Not only that, but the system independence encourages a collaborative work environment in which many users, rather than a single user, are in control, allowing for the more accurate detection of human gestures of various sizes and colors. Remember that each of these technological enablers has advantages and disadvantages. The requirement of physical contact for contact-enabled devices can be discomforting to the users; however, such devices provide high recognition accuracy with reduced complexity in terms of implementation. As much as vision-enabled devices are user-friendly, they face configuration complexity in addition to occlusion.

1.4. Contribution

This article reviews current developments in the field of human–computer interaction (HCI). The emphasis is on the different application areas, where hand gestures are used to create effective engagement. The goal of this article is to give an overview of the current status of static and dynamic hand gesture recognition in the area of HCI, including gesture taxonomies, representations, and recognition methods, as well as to identify future research objectives in the field.

1.5. Research Questions

The following major questions are addressed by this research:

1. What are the main difficulties faced in gesture recognition?
2. What are some challenges faced with gesture recognition?
3. What are the major algorithms involved in gesture recognition?

1.6. Organization of the Work

The rest of the paper is organized as follows: Section 3 introduces the types of hand gestures. Section 4 considers an evaluation of the latest research works in regard to hand gesture recognition. Section 5 considers the current research challenges in this field. Lastly, Section 5 provides the conclusion. The overall organization of the work is nicely depicted in Figure 4.

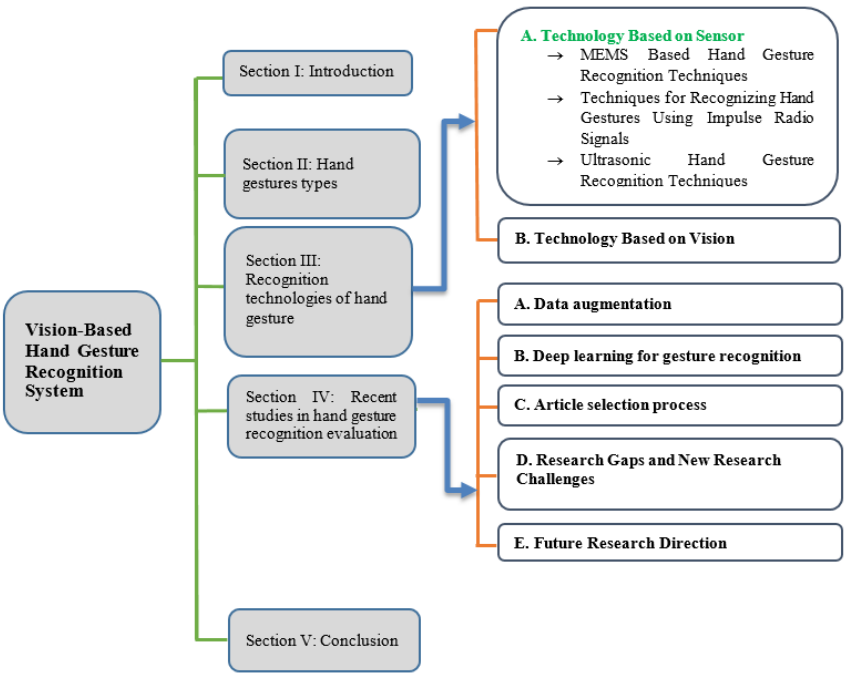

Figure 4. Five class graphical presentation.

2. Hand Gestures Types

Hand gestures are a kind of body language in which the position and shape of the center of the palm and the fingers communicate specific information. The gesture is made up of both static and dynamic hand movements in general. Dynamic hand gestures are made up of a series of hand movements, while static hand gestures are based only on the shape of the hand. Different individuals describe gestures differently due to the cultural variety and uniqueness of gestures. Static hand gestures rely on the shape of the hand gesture to convey the message, while dynamic hand gestures rely on the movement of the hands to transfer the meaning. The ability to instantly and without delay identify hand motions is known as the detection of real-time hand gestures. Processing speed, image processing techniques, acceptable delay in conveying results, and recognition algorithms differ between real-time and non-real-time hand gestures.

3. Recognition Technologies of Hand Gesture

The results of the research show that hand gestures allow technology to be divided into three types: sensor-driven, vision-driven and deep learning. Sensor-based technology, as its name suggests, uses different sensors such as the accelerometer and the gyroscope, while RGB cameras and infrared sensors are used to extract and identify properties from a collection of datasets of hand movements, respectively. The general framework about hand gesture recognition and standard framework for hand gesture recognition using Kinect is shown in Figures 5 and 6, respectively.

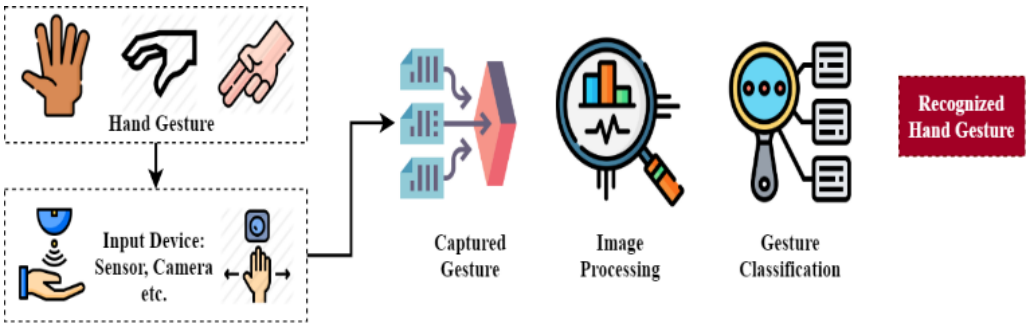

Figure 5. General framework about hand gesture recognition.

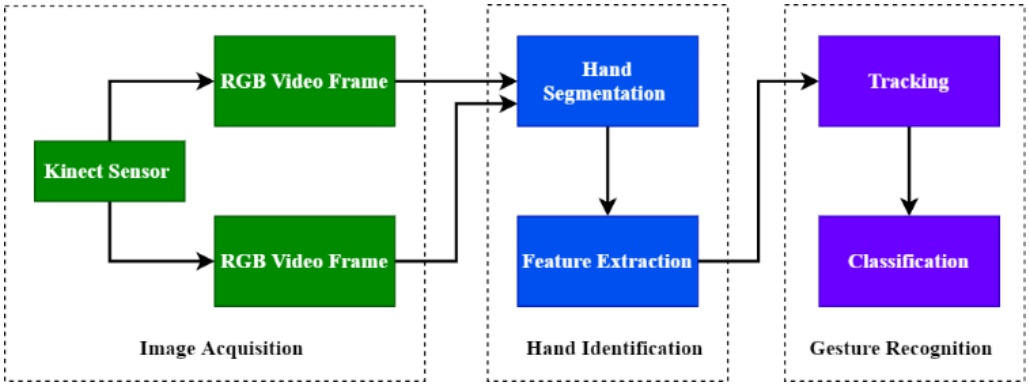

Figure 6. Standard framework for hand gesture recognition using Kinect.

3.1. Technology Based on Sensor

In sensor-based hand gesture recognition algorithms, motion sensors, which are either integrated into gloves or utilized in smart devices such as smartphones using built-in accelerometers with gyroscope sensors, are used.

The main objective is to gather and use triaxial data for the proper application. This method of gesture recognition in a smart gadget that uses a three-axis accelerometer and Gyro sensor, Ref. [1] presents a continuous hand gesture identification (CHG)

technique to constantly identify hand movements. A Samsung ATIV smartphone is employed as an intelligent tool for gesture motions, and smart devices are operated using the Samsung AllShare protocol. Since the waveforms reflect change in amplitude and phase, the machine techniques such as CNN may be used for sensing them. Ref. [2] presented the feedforward neural network and similarity matching (FNN/SM) hand motion detection technique for accelerometer-based pen type sensing devices. A triaxal accelerometer, which is preprocessed by a moving average filter, is used to identify hand motion acceleration data. There is also a segmentation algorithm to control each key motion on the fly at the initial and final positions. Using the basic gesture samples used to train the FNN model after feature extraction, the fundamental gesture sequences are categorized. The series of basic acts is then coded using the codes of Johnson. Finally, the complex gesture is found by comparing the anticipated basic series of gestures with frequent template sequences.

3.1.1. Techniques for Recognizing Hand Gestures Using Impulse Radio Signals

The transmission (Tx) produces an infrared signal through its antenna and sends it. The waved shape of the hand (Rx), consisting of an antenna, amplifiers, a low-pass filter and an oscilloscope with a high speed, is received from the receiver. Due to the varying amplitude and phase of the reflected waveforms, movement shapes may be used in machine learning methods, such as CNN. It is demonstrated in the picture below that this technique works.

3.1.2. Ultrasonic Hand Gesture Recognition Techniques

In this technique, loudspeakers and microphones are utilized as ultrasonic I/O devices. The Doppler shift of ultrasonic waves reflected in a moving human body is used in this technique. During a gesture, the system constantly samples the ultrasound. It produces a series of time variations which are rich in the distinct characteristics of each action. We classify future motions using a mixture of fundamental patterns and supervised methods of machine learning.

3.2. Technology Based on Vision

The three essential phases of a vision-based method are image acquisition, image segmentation and lastly image identification. Many academics have developed hand motion detection systems based on these three stages in real time. The dynamic recognition of hand gestures is the identification of a moving hand with a number of motions, whereas the recognition of the hand position is static hand gestures. Hand motions may be visually categorized utilizing methods based on 3D modeling and looks. Those are the visual sub models of 3D models and model-based appearance methods. The 3D hand gesture model provides a 3D spatial representation of the human hand with time automation. The four kinds of appearance-based hand gesture representation methods are color, silhouette, decorative, and motion-based models.

Vision-based techniques for recognizing gestures: as previously mentioned, there are three basic stages: detection, tracking, and recognition. Ref. [3] explains briefly the numerous sub-techniques employed at various phases of their study. The main phases in the detecting phase include color, shape pixel, 3D model, and motion. Correlation-based and contour-based tracking are two kinds of template-based tracking. Other tracking techniques include optimum estimation, particle filtering, and camshift. Using a range of algorithms and machine learning techniques, the last stage in the complex identification process is to identify static and dynamic hand movements. Time delay neural networks, hidden Markov models, dynamic time warping networks, and finite state machines are examples of these techniques.

The most frequent use of color detection is the detection of skin color on the hand. The color space conversion to either HSV or YCbCr is done first for reliable hand segmentation. Binarization is achieved using skin color threshold values, and noise is reduced using image processing techniques, such as morphological processing. The segmented

hand is then identified using a variety of techniques for extracting features and recognizing the segmented hand. One of the study articles by [4] was based on robust hand motion segmentation utilizing YCbCr color space and K-means clustering.

Tracking is essential since it sees and detects the hand in real time for the identification of hand gestures. Many tracking methods, one based on contour moments, were proposed. In order to perform hand center identification, ref. [5] utilized transformation and contour detection on complicated backdrops, which could be used for hand tracking. After that, the fingertip position method used to detect hand motions was calculated using a convex hull. In 1972, Sklansky launched the three-coin technique of the convex hull.

4. Significant Research Works on Hand Gesture Recognition

The focus in the preceding part was on the work's initial components. This section builds on that basis by providing in-depth coverage of recent hand gesture recognition research studies that have been selected and reviewed. These works were carefully chosen and adapted to satisfy the challenging requirements in this field. Some of the works analyzed were clearly not evaluated in previous publications, implying that they are recent works. In the case of identity-based hand gestures, ref. [6] used a multivariate Gaussian distribution. The scientists divided the image into two procedures in their study: skin color-based segmentation and cluster-based thresholding. The restructuring of sign language is accomplished using a variety of approaches. Ref. [7] proposed a random forest-based wearable motion recording sensor for hand gesture detection. Gesture characteristics were gathered at various time intervals to quantify hand movements. The experimental evaluation is based on a dataset of gestures from a gesture pool. Ref. [8] enhanced gesture detection by employing the custom classifier in the frictionless operating room. A computer-aided surgical technique was proposed for the categorization of user movements. To forecast the result, vector support machines (SVM) and classifiers from naive Bayes were used.

Hand gesture recognition for healthcare was developed by [9]. To identify the hand motions without any noise, a convolutional neural network (CNN) was suggested. For segmentation, a MobiGesture was required to enhance accuracy and recall depending on user motions. Ref. [10] created a wearable sensor with a double surface EMG. Support vector machines were used to classify gesture recognition (SVM). Using two EMG channels on flexors and extensors, four types of gestures were identified.

The project's purpose was to eliminate the gesture's failure tolerance. An EMG armband was created by [11] to recognize human motions based on physiological characteristics. Wearing-independent hand gesture recognition has been developed. The EMG elements gave the gesture recognition a distinct scale. The uniform signal helped to enhance the recognition. A random forest was utilized to determine how gestures would progress and be pursued. Ref. [12] investigated dynamic hand movements using spatial–temporal algorithm convolutional networks. Three types of graph edges associated with the activity of hand joints were proposed in a skeleton-based model. A deep neural network was utilized to pick semantic characteristics in order to provide an accurate output. Ref. [13] created a string matching technique for understanding hand motions in real-time situations. The k-means technique was used to create an approximation string matching to capture the features of hand joints. It was done to enhance the precision of various motions by specifying the amount of clusters. Ref. [14] discovered maneuvers using the convolutional pose machine (CPM) and Gaussian mixing models of Zhang et al. (FGMM). The first stage was to acquire critical hand points with the CPM. The FGMM was then used to filter and categorize non-gestures by critical point. Ref. [15] developed a real-time fine gesture detection system using a convolution neural network. The system for monitoring muscle contraction in the forearms was built using the frequency–time–space cross-domain pre-processing method. Ref. [16] used jump motion with spatial fuzzy matching (SFM) to enhance hand gesture detection. The gesture matching was performed by analyzing

the fused gesture dataset, where the gesture frames were categorized. The SFM was then utilized to accelerate the analysis processing.

To improve the efficiency of gesture analysis, ref. [17] created online dynamic hand gesture detection. The technique was selected for three reasons: there was no evidence of when the gesture began and ended; the rearrangement was only recognized once; and the gesture under investigation had to be within memory and budget. Gesture recognition and categorization are aided by the use of CNN to build a two-level hierarchy structure (TLHS).

Ref. [18] used long short-term memory(LSTM) to accomplish continuous hand gesture recognition. The analysis was used to detect input gestures using accelerators and/or gyroscopes. The output was acquired using LSTM, which was accomplished by accurately categorizing the generated gesture. Ref. [19] suggested a hand gesture approach to control powerpoint presentation. Primarily, the proposed approach can ensure the storage of gesture images without need for a database.

To up-scale the mouse capabilities, ref. [20] proposed a method based on the mouse pointing device. In addition to being a vision-based interface, the technique uses a camera. The solution can recognize a variety of predefined hand postures, including gesture mode and mouse mode revealing. The method uses the HSV color space to identify skin after obtaining the pictures.

A technique for guaranteeing the identification of gestures in real time was devised by [21]. A programmable field gate array was used to develop the technique. Ten distinct static hand movements may be recognized by the program. To represent the system, the authors used the verilog hardware description. Initially, the picture is acquired using a camera with a complementary metal oxide semiconductor image sensor during implementation. The image preparation is then completed. Finally, the system uses the YCbCr color space to conduct skin color segmentation on that picture. The 25 unique hand pictures are evaluated for gesture categorization, resulting in a 94.40 percent recognition rate. Ref. [22] examined the human–computer interaction from the standpoint of hand gesture recognition. The authors discussed past hand gesture detection efforts as well as the technological challenges they encountered. Vision-driven, glove-driven, and depth-driven techniques are all being explored. Finally, the study analyzes Microsoft's release of Kinect data for finger and hand motion recognition.

To operate Power point and VLC media players, ref. [23] proposed a hand gesture detection method. Ref. [24] described vision studies conducted over the past 15 years that focused on identifying hand gesture methods. In addition, this study covers more than two dozen open-access hand gesture datasets, as well as download links. Ref. [25] proposed a framework for hand gesture recognition. The scientists looked at how they might utilize data from Kinect to recover depth information on all of the pixels in a image. They also developed a method for identifying the corresponding closed contour points in the desired depth intervals. The palm center is recognized right away from the recovery points. Furthermore, the fingers were detected using the k-curvature method. The first in, first out (FIFO) method was used to gather the contour points, palm center, and fingers throughout the procedure. After that, a DTW algorithm was used to identify gestures. Ref. [26] proposed the use of vision technologies to detect hand movements. To obtain the required outcomes, the authors used techniques, such as skin color filtering, edge detection, and convex hull. Ref. [27] looked at a technique for identifying hand gestures that utilizes 3D depth sensors. Three-dimensional hand modeling, static hand motion, and hand route gesture are among the work fields. The authors emphasized gesture recognition techniques in their work, as well as the regions that are utilized in such systems. Ref. [28] used a vision-driven recognition technique, as well as a library of hand gestures based on skin color architecture and threshold approaches through PCA template matching. The procedure starts with hand segmentation, which is done using a skin color model. The hand picture is then separated from the backdrop using the Otsu thresholding technique. Finally, the PCA employs a template-driven matching approach to accomplish gesture recognition. Ref. [29] considered a review of the Kinect-based markerless two-hand gesture detection

method. The authors utilize Kinect's depth and skeleton to accomplish marker-free hand extraction. The finger Earth mover's distance (FEMD) is utilized for gesture measurement in the novel morphological finger extraction technique. Furthermore, for natural and reliable human–computer interaction, two hand-driven user interfaces are utilized. The collected results demonstrate the system's efficacy.

Gesture communication was studied by [30]. Their research highlighted the importance of both nonverbal and vocal communication. Furthermore, they utilized nonverbal communication in their 3D game. Four stages are explored while developing a hand gesture recognition system. Data collection, segmentation, feature extraction, and gesture recognition are among these stages.

In this research, ref. [31] examined the importance of delivering a message from a user to a person through a man–machine interface. In the vision-driven approach, their work disregards and eliminates the usage of sensors linked to the human body. Furthermore, two important apps are used to operate the TV and mobile device via hand gestures. Through the use of gestures, the user can ably achieve direct communication with the system. The solution focused on the need to ensure successful user-friendly, as well as machine-friendly applications. The hand gesture selection module enables the user to customize the interface of TV control through selection of the shapes and hand motions.

Ref. [32] proposed an RGB-D sensor-based hand gesture detection method. The method utilizes deep knowledge to minimize light conditions and overwhelming background problems. The four stages of the method are hand segmentation, extraction of characteristics, static gesture classification and dynamic gesture classification. Segmentation of the hand is also carried out by segmenting the skin color, while background removal is utilized to segment the hand. The hand is removed afterwards. While fingertip detection is used to create static movements, dynamic gestures occur at the Euclidean distance. All this consists of 90 photos of 9 individuals with 10 dynamics and 6 motions. Ref. [33] proposed a hand gesture-based HCI approach. The system is built on vision-based technology and uses machine learning methods to accomplish its goals. Pre-processing, feature extraction, and recognition are the three stages of the technique. The system uses an ML classifier to achieve recognition. SVM is used to deal with static motions, whereas HMM is used to deal with dynamic gestures. The experimental findings indicate that employing 11 pre-defined motions, SVM and HMM achieved an accuracy rating of 99.7% and 93.7%, respectively.

For HCI, ref. [34] proposed a hand gesture detection system based on vision techniques. The authors developed a real-time application that uses a color-based detection method instead of ANN training to restrict mouse movement inside windows. Ref. [35] proposed a method for determining finger and hand posture based on depth information. To solve the problem of merging fingers, the method shows an apex-shaped hand structural contour. To accomplish hand identification, the authors also use depth thresholding. Using a dataset of 1000 postures, the findings show a 99.1% accuracy with FD and a 96.3 percent accuracy with global finger information. Ref. [36] suggested a depth sensor home-based hand gesture recognition method. The method is divided into two stages. The first step allows the creation of the required database. In the second step, the system extracts various features from the labeled hand parts to issue commands to the devices. Ref. [37] suggested a shape parameter based method. Moreover, to generate the shape information, a computer vision method was employed. Basically, the proposed hand gesture solution is matched through shape-enabled techniques. Ref. [38] recommended a novel approach for hand part segmentation based on the use of adaptive skin color architecture. Initially, the method grabs hand part pixel values and background part, thus changing them into the YCbCr color space architecture. Next, skin and Gaussian models are presented. Lastly, the approach performs segmentation. Ref. [39] used image and computer vision technologies to accomplish computer-aided control and monitor utilizing hand motions. Furthermore, the Haar classifier is used to guarantee face detection, and YCbCr is used to achieve hand gesture recognition. Hand tracking is done via the hand tracking technique, and features are extracted using the convexity defect hull. Furthermore, the hand region

feature is controlled by the mouse position. Ref. [40] looked at how deep estimation and gesture recognition might be used. In terms of suggested techniques for hand recognition and gesture classifications, the authors focused on depth-based gesture identification. In the work, 13 methods, 11 of which were determined to be related to hand localization and gesture classification, in that order. Moreover, both the Kinect sensor and OpenNI were employed to ensure hand tracking. In total, 37 works were covered pertaining to the definition of gesture type and classification based on cluster depth. Ref. [24] conducted a review on the roles of the HCI system in the successful realization of hand gesture recognition. Their work shows that for current research, issues in hand gesture recognition include sensitivity in regard to size, shape speed variability and occlusion concerns. Their work focused on algorithms for vision-based hand gesture recognition. Moreover, they compared both quantitative and qualitative algorithms using RGB and RGB-D cameras. Additionally, several experimental simulation methodologies and measures were employed to evaluate the algorithms. Also reviewed are tens of publicly accessible hand gesture databases and the related links for download. Ref. [41] presents a useful idea for using mouse points to perform a variety of tasks, including single and double clicking, dragging, and so on. They claim that recognizing motions is challenging since many factors are involved, including modeling and movement, analysis, and pattern recognition. Ref. [42] developed a method to evaluate the dissimilarity of hand movements based on a super-pixel earth mover's distance (SP-EMD). Furthermore, the texture and depth of the hand are conveyed via super-pixels, allowing the color of the motion to be recognized. It becomes invariant in terms of scaling, translation, and rotation after appropriate processing. The collected findings show that the identification speed is quick and the mean accuracy is high.

HCI was regarded by [22] to be an essential component of hand gesture recognition. They reviewed the history of hand gesture recognition and the technological challenges that come with it. The authors also explained the different methods, such as vision-based, glove-based, and depth-based ones. In addition, the authors discussed research projects using Microsoft's Kinect device data, which includes novel finger identification and hand motion recognition. Finally, attention is given to Kinect-based applications, such as clinical surgery and robotics. Ref. [27] investigated the field of hand gesture detection using 3D depth sensors. The authors demonstrated that commercial depth sensors and publicly available datasets are widely used in the area of 3D. In addition, the systems' different fields include 3D hand modeling, static gesture, hand trajectory gesture, and continuous hand gesture detection. Furthermore, the system makes use of cutting-edge research for 3D hand gesture detection. The work's primary emphasis is on gesture recognition techniques and a clear description of the areas where this approach is used. Ref. [43] looked at using a vision-based technique to create a natural and intuitive interface. The authors concentrated on approaches for recognizing dynamic hand gestures. Their research is divided into three sections: detection and segmentation, tracking, and classification. The numerous segmentation methods based on skin color, shape, particle filtering, TLD, camshaft, and other factors are explained. In terms of applications, it has had a significant influence on daily life in a variety of ways.

The research of [44] is focused on gesture-based first-person control. The gamepad and a combination of keyboard and mouse are the two methods of input. The end product shows how to operate a first-person shooter game with a single hand gesture, followed by a comparison of standard input techniques [45]. Around 26 people collaborate to give the outcomes, which include summaries of inline performance and a survey of previous games [46]. The player's abilities were developed using their outcomes [25,47]. There are designs in the scientific literature which are designed to identify and classify the hand gestures of static and dynamic types. These ideas are built on infrared pictures, color images and depths [48–50]. This study focuses on a literature review of hand gesture strategies and discusses their pros and limits in various situations. In addition, the performance of these methods is tabulated, with an emphasis on computer vision techniques that deal with similarity and difference points; hand segmentation techniques; classification

algorithms and limitations; number and types of gestures; dataset used; detection range (distance); and camera type [51]. Convolutional neural networks (CNN) are used to categorize images of hand gestures. A newly developed metaheuristic technique, the Harris hawks optimization (HHO) algorithm, is utilized to optimize the CNN's hyperparameters. Their extensive comparison research shows that the proposed HHO-CNN hybrid model outperforms current models by achieving 100 percent accuracy [52].

This article examines the flex, accelerometer, and gyroscope-based smart prototype developed to recognize sign language motions. These sensors are put on a glove in order to record and assemble alphabetic (i.e., 0–10, A–Z) and numeric (i.e., 0–10, and A–Z datasets). The primary purpose of the proposed model is to categorize sign gestures produced by deaf–mute people and identify the true meaning of movements performed [53].

Based on the ultrasonic frequency modulated continuous wave (FMCW) and the ConvLSTM model, a system for gesture identification was suggested in this work. It uses a hardware configuration consisting of one transmitter and three spatially separated receivers [54]. In this research, the authors proposed a hand gesture detection system for a dataset of lowercase numbers and alphabets. The suggested method recognizes the hand using information on skin color and mobility. Hand tracking is performed using a two-level tracking system and a modified Kanade–Lucas–Tomasi (KLT) tracking algorithm [55]. Table 1 presents the comparison of the most recent selected gesture recognition studies.

Table 1. Comparison of the most recent selected gesture recognition studies.

Author	Findings	Challenges
[56]	In this study, image processing techniques such as wavelets and empirical mode decomposition were suggested to extract picture functionalities in order to identify 2D or 3D manual motions. Classification of artificial neural networks (ANN), which was utilized for the training and classification of data in addition to the CNN (CNN).	Three-dimensional gesture disparities were measured utilizing the left and right 3D gesture videos.
[57]	Deaf–mute elderly folk use five distinct hand signals to seek a particular item, such as drink, food, toilet, assistance, and medication. Since older individuals cannot do anything independently, their requests were delivered to their smartphone.	Microsoft Kinect v2 sensor's capability to extract hand movements in real time keeps this study in a restricted area.
[58]	The physical closeness of gestures and voices may be loosened slightly and utilized by individuals with unique abilities. It was always important to explore efficient human computer interaction (HCI) in developing new approaches and methodologies.	Many of the methods encounter difficulties like occlusions, changes in lighting, low resolution and a high frame rate.
[59]	A working prototype is created to perform gestures based on real-time interactions, comprising a wearable gesture detecting device with four pressure sensors and the appropriate computational framework.	The hardware design of the system has to be further simplified to make it more feasible. More research on the balance between system resilience and sensitivity is required.
[60]	This article offers a lightweight model based on the YOLO (You Look Only Once) v3 and the DarkNet-53 neural networks for gesture detection without further preprocessing, filtration of pictures and image improvement. Even in a complicated context the suggested model was very accurate, and even in low resolution image mode motions were effectively identified. Rate of high frame.	The primary challenge of this application for identification of gestures in real time is the classification and recognition of gestures. Hand recognition is a method used by several algorithms and ideas of diverse approaches for understanding the movement of a hand, such as picture and neural networks.
[61]	This work formulates the recognition of gestures as an irregular issue of sequence identification and aims to capture long-run spatial correlations in points of the cloud. In order to spread information from past to future while maintaining its spatial structure, a new and effective PointLSTM is suggested.	The underlying geometric structure and distance information for the object surfaces are accurately described in dot clouds as compared with RGB data, which offer additional indicators of gesture identification.

Table 1. *Cont.*

Author	Findings	Challenges
[62]	A new system is presented for a dynamic recognition of hand gestures utilizing various architectures to learn how to partition hands, local and global features and globalization and recognition features of the sequence.	To create an efficient system for recognition, hand segmentation, local representation of hand forms, global corporate configuration, and gesture sequence modeling need to be addressed.
[63]	This article detects and recognizes the gestures of the human hand using the method to classification for neural networks (CNN). This process flow includes hand area segmentation using mask image, finger segmentation, segmented finger image normalization and CNN classification finger identification.	SVM and the naive Bayes classification were used to recognize the conventional gesture technique and needed a large number of data for the identification of gesture patterns.
[64]	They provided a study of existing deep learning methodologies for action and gesture detection in picture sequences, as well as a taxonomy that outlines key components of deep learning for both tasks.	They looked through the suggested architectures, fusion methodologies, primary datasets, and competitions in depth. They described and analyzed the key works presented so far, focusing on how they deal with the temporal component of data and suggesting potential and challenges for future study.
[65]	They solve the problems by employing an end-to-end learning recurrent 3D convolutional neural network. They created a spatiotemporal transformer module with recurrent connections between surrounding time slices that can dynamically change a 3D feature map into a canonical view in both space and time.	The main challenge in egocentric vision gesture detection is the global camera motion created by the device wearer's spontaneous head movement.
[66]	To categorize video sequences of hand motions, a long-term recurrent convolution network is utilized. Long-term recurrent convolution is the most common kind of long-term recurrent convolution. Multiple frames captured from a video sequence are fed into a network to conduct categorization in a network-based action classifier.	Apart from lowering the accuracy of the classifier, the inclusion of several frames increases the computing complexity of the system.
[67]	The MEMP network's major characteristic is that it extracts and predicts the temporal and spatial feature information of gesture video numerous times, allowing for great accuracy. MEMP stands for multiple extraction and multiple prediction.	They present a neural network with an alternative fusion of 3D CNN and ConvLSTM since each kind of neural network structure has its own constraints. MEMP was developed by them.
[68]	This research introduces a new machine learning architecture that is especially built for gesture identification based on radio frequency. They are particularly interested in high-frequency (60 GHz) short-range radar sensing, such as Google's Soli sensor.	The signal has certain unique characteristics, such as the ability to resolve motion at a very fine level and the ability to segment in range and velocity space rather than picture space. This allows for the identification of new sorts of inputs, but it makes the design of input recognition algorithms much more challenging.
[69]	They propose learning spatio-temporal properties from successive video frames using a 3D convolutional neural network (CNN). They test their method using recordings of robot-assisted suturing on a bench-top model from the JIGSAWS dataset, which is freely accessible.	Recognizing surgical gestures automatically is an important step in gaining a complete grasp of surgical expertise. Automatic skill evaluation, intra-operative monitoring of essential surgical processes, and semi-automation of surgical activities are all possible applications.
[70,71]	They blur the image frames from videos to remove the background noise. The photos are then converted to HSV color mode. They transform the picture to black-and-white format through dilation, erosion, filtering, and thresholding. Finally, hand movements are identified using SVM.	Gesture-based technology may assist the handicapped, as well as the general public, to maintain their safety and requirements. Due to the significant changeability of the properties of each motion with regard to various persons, gesture detection from video streams is a complicated matter.

Table 1. *Cont.*

Author	Findings	Challenges
[72,73]	The purpose of this study is to offer a method for Hajj applications that is based on a convolutional neural network model. They also created a technique for counting and then assessing crowd density. The model employs an architecture that recognizes each individual in the crowd, marks their head position with a bounding box, and counts them in their own unique dataset (HAJJ-Crowd).	There has been a growth in interest in the improvement of video analytics and visual monitoring to better the safety and security of pilgrims while in Makkah. It is mostly due to the fact that Hajj is a one-of-a-kind event with hundreds of thousands of people crowded into a small area.
[74]	This study presents crowd density analysis using machine learning. The primary goal of this model is to find the best machine learning method for crowd density categorization with the greatest performance.	Crowd control is essential for ensuring crowd safety. Crowd monitoring is an efficient method of observing, controlling, and comprehending crowd behavior.

4.1. Data Augmentation

Data augmentation is a technique of expanding the data set by producing various picture shapes to increase model performance [75]. It also helps to mitigate the over-fitting issue in the model during the training stage. The overcast issue arises when random noise or mistakes occur instead of when the underlying connection is there. Using an increase in data, additional images were produced for the model from each picture because some irrelevant patterns may occur throughout the model training process. Several methods were employed for data augmentation operations: rotational changes, vertical and horizontal rotations, and intensity disorder, including light disturbances [76–79].

4.2. Deep Learning for Gesture Recognition

A classical ANN involves a local minimal issue, which typically ends with a local optimization process rather than a globally optimal state. More overfitting issues often complicates general machine learning models. Intensive network structure optimization may address the issues of the local minima and overriding by DNNs [80,81]. Deep learning is a machine learning-based approach that educates computers to accomplish tasks similar to those performed by humans. For example, deep learning is the underlying technology that enables driverless automobiles to detect traffic lights and people. It is also the underlying principle of audio and speech recognition in a variety of devices, such as mobile phones and tablets. Deep learning is gaining popularity because it is capable of performing tasks that were previously impossible. A deep learning model is constructed by layering data, which may be images, text, or audio, into distinct and discrete categorization layers. Artificial intelligence has the potential to provide findings that are 100 percent accurate with human-level precision and potentially beat humans in terms of speed. These models are developed using big data sets and machine learning approaches, such as CNN or ANN, both of which include several categorization layers. In machine learning approaches, the system would instruct the user on how to correctly utilize the model via the use of pictures, voice, and text. Deep learning models provide precise, accurate outcomes that are on par with human performance. These models are constructed using the data provided and turn them into artificial neural networks with many layers of categorized data. General data following the diagram of a fully convolutional neural network (FCNN) for gesture recognition are depicted below in Figure 7. Deep learning achieves unprecedented precision and accuracy, which enables it to match customer expectations. It is beneficial in a variety of applications, including autonomous vehicles. Recent advancements have shown that artificial intelligence is capable of outperforming humans in classifying photos. Deep learning requires a vast volume of labeled data. For example, building driverless vehicles requires the collection of hundreds of thousands of photos and movies. Deep learning needs an enormous amount of computational power. Elite GPUs are comparable in architecture and are well suited for deep learning. When cloud computing and clusters are integrated, it takes much less time than it did before, when it took weeks. Due to the

fact that deep learning is based on neural networks, it is often referred to as "deep neural networks". The term "deep" is often used to refer to the number of hidden layers in the neural system. Typically, neural networks comprise just 2–3 hidden layers; however, deep systems may contain up to 150 layers. They must train the deep learning models prior to implementing them. They need a vast quantity of labeled data and neural networks to train these models. This enables them to derive the characteristics directly from the data without requiring any human input.

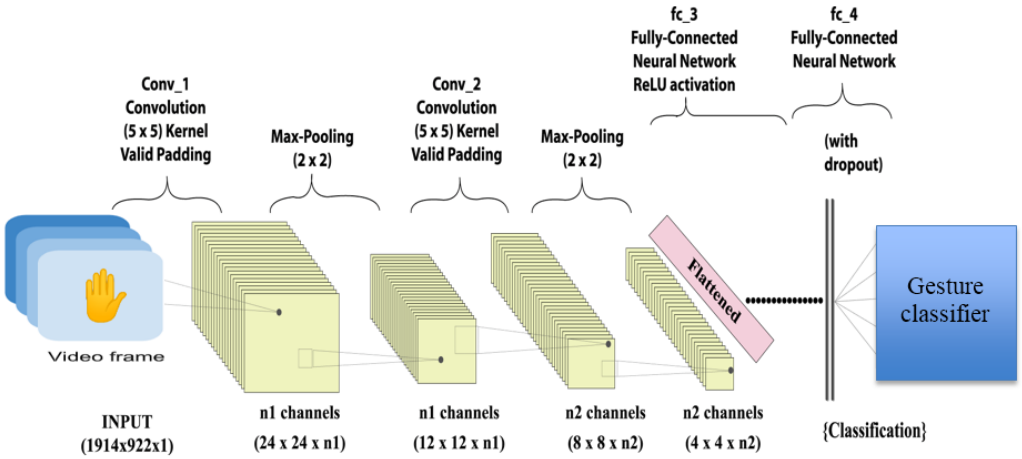

Figure 7. Gesture recognition using FCNN.

One of the most well-known deep neural network methods is CNN. "Conventional neural network" is the abbreviation for "conventional neural network". It employs 2D convolutional layers to handle 2D data and uses categorized layers of input data [82,83].

4.3. Summary

We give a review of current convolutional neural networks for action and gesture identification in image frames in this study. We provide a framework for handling both problems that covers key elements of deep learning as well as other hand-crafted methods. The suggested architectures, fusion methodologies, primary datasets, and competitions are discussed in depth. We outline and examine the key suggestions thus far, with a focus on how they deal with the data temporal component, suggesting potential and problems for future study based on 3D models [84–92]. Based on the evaluation, the associated difficulties and future research directions were discussed in the preceding section. It is critical to develop practical answers in the future. We think that the conversations in this section of the work will reveal fresh research gaps that will help us get closer to the much-desired next-generation technologies [93–97]. Hybrid models that combine handcrafted and new descriptors are predicted to advance. Similarly, we believe that deep learning solutions for large-scale, real-time action and gesture identification would be of interest to the community. In extended, uncut, and realistic videos, immediate effort is also anticipated in action/gesture localization. As a result, we anticipate that in the next years, emerging challenges, such as early recognition, multi-task learning, captioning, recognition from low resolution sequences, and life log devices will attract attention [97–102].

5. Conclusions

This paper provides a systematic review and analysis of recent vision-based gesture recognition methods in the design of more efficient and intelligent HCIs. In the area of vision-based hand gesture recognition, significant development has been achieved in the recent few years, both in terms of hardware and software. The evaluation results also show that the identification of hand gestures within the scientific community has created a great deal of attention among a broad range of techniques for recognizing vision-based

gestures. Over the course of 11 years, this article looked at the difficulties and development of the vision-based hand gesture recognition system. Data gathering, features, and training environment seem to be covered in almost every article we investigated.

We have discussed the difficulties and challenges associated with gesture recognition in this article. We reviewed the most important algorithms used in gesture recognition as well. Hand gesture recognition is anticipated to play a significant part in our everyday lives in the modern world. The surrounding gadgets will almost certainly all have hand gesture interfaces sooner than one may imagine. The recognition of hand gestures is anticipated to play an important part in our daily lives. In the modern world, most of the technologies around us are mostly controlled by hand gestures. In the future, we want to improve our analytical approach to learn more about gesture recognition techniques.

Funding: This work was supported by Multimedia University, Cyberjaya, Selangor, Malaysia. The grant number is MMUE/210029.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: There is no conflict of interest in this research.

References

1. Gupta, H.P.; Chudgar, H.S.; Mukherjee, S.; Dutta, T.; Sharma, K. A continuous hand gestures recognition technique for human-machine interaction using accelerometer and gyroscope sensors. *IEEE Sens. J.* **2016**, *16*, 6425–6432. [CrossRef]
2. Xie, R.; Cao, J. Accelerometer-based hand gesture recognition by neural network and similarity matching. *IEEE Sens. J.* **2016**, *16*, 4537–4545. [CrossRef]
3. Rautaray, S.S.; Agrawal, A. Vision based hand gesture recognition for human computer interaction: A survey. *Artif. Intell. Rev.* **2015**, *43*, 1–54. [CrossRef]
4. Zhang, Q.-Y.; Lu, J.-C.; Zhang, M.-Y.; Duan, H.-X. Hand gesture segmentation method based on YCbCr color space and K-means clustering. *Int. J. Signal Process. Image Process. Pattern Recognit.* **2015**, *8*, 105–116. [CrossRef]
5. Lai, H.Y.; Lai, H.J. Real-time dynamic hand gesture recognition. In Proceedings of the 2014 International Symposium on Computer, Consumer and Control, Taichung, Taiwan, 10–12 June 2014; pp. 658–661.
6. Hasan, M.M.; Mishra, P.K. Features fitting using multivariate gaussian distribution for hand gesture recognition. *Int. J. Comput. Sci. Emerg. Technol. Ijcset* **2012**, *3*, 73–80.
7. Bargellesi, N.; Carletti, M.; Cenedese, A.; Susto, G.A.; Terzi, M. A random forest-based approach for hand gesture recognition with wireless wearable motion capture sensors. *IFAC-PapersOnLine* **2019**, *52*, 128–133. [CrossRef]
8. Cho, Y.; Lee, A.; Park, J.; Ko, B.; Kim, N. Enhancement of gesture recognition for contactless interface using a personalized classifier in the operating room. *Comput. Methods Programs Biomed.* **2018**, *161*, 39–44. [CrossRef]
9. Zhao, H.; Ma, Y.; Wang, S.; Watson, A.; Zhou, G. MobiGesture: Mobility-aware hand gesture recognition for healthcare. *Smart Health* **2018**, *9*, 129–143. [CrossRef]
10. Tavakoli, M.; Benussi, C.; Lopes, P.A.; Osorio, L.B.; de Almeida, A.T. Robust hand gesture recognition with a double channel surface EMG wearable armband and SVM classifier. *Biomed. Signal Process. Control.* **2018**, *46*, 121–130. [CrossRef]
11. Zhang, Y.; Chen, Y.; Yu, H.; Yang, X.; Lu, W.; Liu, H. Wearing-independent hand gesture recognition method based on EMG armband. *Pers. Ubiquitous Comput.* **2018**, *22*, 511–524. [CrossRef]
12. Li, Y.; He, Z.; Ye, X.; He, Z.; Han, K. Spatial temporal graph convolutional networks for skeleton-based dynamic hand gesture recognition. *Eurasip J. Image Video Process.* **2019**, *2019*, 78. [CrossRef]
13. Alonso, D.G.; Teyseyre, A.; Soria, A.; Berdun, L. Hand gesture recognition in real world scenarios using approximate string matching. *Multimed. Tools Appl.* **2020**, *79*, 20773–20794. [CrossRef]
14. Zhang, T.; Lin, H.; Ju, Z.; Yang, C. Hand Gesture recognition in complex background based on convolutional pose machine and fuzzy Gaussian mixture models. *Int. J. Fuzzy Syst.* **2020**, *22*, 1330–1341. [CrossRef]
15. Tam, S.; Boukadoum, M.; Campeau-Lecours, A.; Gosselin, B. A fully embedded adaptive real-time hand gesture classifier leveraging HD-sEMG and deep learning. *IEEE Trans. Biomed. Circuits Syst.* **2019**, *14*, 232–243. [CrossRef]
16. Li, H.; Wu, L.; Wang, H.; Han, C.; Quan, W.; Zhao, J. Hand gesture recognition enhancement based on spatial fuzzy matching in leap motion. *IEEE Trans. Ind. Inform.* **2019**, *16*, 1885–1894. [CrossRef]
17. Köpüklü, O.; Gunduz, A.; Kose, N.; Rigoll, G. Online dynamic hand gesture recognition including efficiency analysis. *IEEE Trans. Biom. Behav. Identity Sci.* **2020**, *2*, 85–97. [CrossRef]
18. Tai, T.M.; Jhang, Y.J.; Liao, Z.W.; Teng, K.C.; Hwang, W.J. Sensor-based continuous hand gesture recognition by long short-term memory. *IEEE Sens. Lett.* **2018**, *2*, 1–4. [CrossRef]

19. Ram Rajesh, J.; Sudharshan, R.; Nagarjunan, D.; Aarthi, R. Remotely controlled PowerPoint presentation navigation using hand gestures. In Proceedings of the International conference on Advances in Computer, Electronics and Electrical Engineering, Vijayawada, India, 22 July 2012.
20. Czupryna, M.; Kawulok, M. Real-time vision pointer interface. In Proceedings of the ELMAR-2012, Zadar, Croatia, 12–14 September 2012; pp. 49–52.
21. Gupta, A.; Sehrawat, V.K.; Khosla, M. FPGA based real time human hand gesture recognition system. *Procedia Technol.* **2012**, *6*, 98–107. [CrossRef]
22. Chen, L.; Wang, F.; Deng, H.; Ji, K. A survey on hand gesture recognition. In Proceedings of the 2013 International Conference on Computer Sciences and Applications, Wuhan, China, 14–15 December 2013; pp. 313–316.
23. Jalab, H.A.; Omer, H.K. Human computer interface using hand gesture recognition based on neural network. In Proceedings of the 2015 5th National Symposium on Information Technology: Towards New Smart World (NSITNSW), Riyadh, Saudi Arabia, 17–19 February 2015; pp. 1–6.
24. Pisharady, P.K.; Saerbeck, M. Recent methods and databases in vision-based hand gesture recognition: A review. *Comput. Vis. Image Underst.* **2015**, *141*, 152–165. [CrossRef]
25. Plouffe, G.; Cretu, A.M. Static and dynamic hand gesture recognition in depth data using dynamic time warping. *IEEE Trans. Instrum. Meas.* **2015**, *65*, 305–316. [CrossRef]
26. Rios-Soria, D.J.; Schaeffer, S.E.; Garza-Villarreal, S.E. Hand-gesture recognition using computer-vision techniques. In Proceedings of the 21st International Conference on Computer Graphics, Visualization and Computer Vision, Plzen, Czech Republic, 24–27 June 2013.
27. Cheng, H.; Yang, L.; Liu, Z. Survey on 3D hand gesture recognition. *IEEE Trans. Circuits Syst. Video Technol.* **2015**, *26*, 1659–1673. [CrossRef]
28. Ahuja, M.K.; Singh, A. Static vision based Hand Gesture recognition using principal component analysis. In Proceedings of the 2015 IEEE 3rd International Conference on MOOCs, Innovation and Technology in Education (MITE), Amritsar, India, 1–2 October 2015; pp. 402–406.
29. Kaur, H.; Rani, J. A review: Study of various techniques of Hand gesture recognition. In Proceedings of the 2016 IEEE 1st International Conference on Power Electronics, Intelligent Control and Energy Systems (ICPEICES), Delhi, India, 4–6 July 2016, pp. 1–5.
30. Sonkusare, J.S.; Chopade, N.B.; Sor, R.; Tade, S.L. A review on hand gesture recognition system. In Proceedings of the 2015 International Conference on Computing Communication Control and Automation, Pune, India, 26–27 February 2015; pp. 790–794.
31. Shimada, A.; Yamashita, T.; Taniguchi, R.I. Hand gesture based TV control system—Towards both user-& machine-friendly gesture applications. In Proceedings of the 19th Korea-Japan Joint Workshop on Frontiers of Computer Vision, Incheon, Korea, 30 January–1 February 2013; pp. 121–126.
32. Palacios, J.M.; Sagüés, C.; Montijano, E.; Llorente, S. Human-computer interaction based on hand gestures using RGB-D sensors. *Sensors* **2013**, *13*, 11842–11860. [CrossRef] [PubMed]
33. Trigueiros, P.; Ribeiro, F.; Reis, L.P. Generic system for human-computer gesture interaction. In Proceedings of the 2014 IEEE International Conference on Autonomous Robot Systems and Competitions (ICARSC), Espinho, Portugal, 14–15 May 2014; pp. 175–180.
34. Dhule, C.; Nagrare, T. Computer vision based human-computer interaction using color detection techniques. In Proceedings of the 2014 Fourth International Conference on Communication Systems and Network Technologies, Washington, DC, USA, 7–9 April 2014; pp. 934–938.
35. Poularakis, S.; Katsavounidis, I. Finger detection and hand posture recognition based on depth information. In Proceedings of the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Florence, Italy, 4–9 May 2014; pp. 4329–4333.
36. Dinh, D.L.; Kim, J.T.; Kim, T.S. Hand gesture recognition and interface via a depth imaging sensor for smart home appliances. *Energy Procedia* **2014**, *62*, 576–582. [CrossRef]
37. Panwar, M. Hand gesture recognition based on shape parameters. In Proceedings of the 2012 International Conference on Computing, Communication and Applications, Dindigul, India, 22–24 February 2012; pp. 1–6.
38. Wang, W.; Pan, J. Hand segmentation using skin color and background information. In Proceedings of the 2012 International Conference on Machine Learning and Cybernetics, Xi'an, China, 15–17 July 2012, Volume 4; pp. 1487–1492.
39. Doğan, R.Ö.; Köse, C. Computer monitoring and control with hand movements. In Proceedings of the 2014 22nd Signal Processing and Communications Applications Conference (SIU), Trabzon, Turkey, 23–25 April 2014; pp. 2110–2113.
40. Suarez, J.; Murphy, R.R. Hand gesture recognition with depth images: A review. In Proceedings of the 2012 IEEE RO-MAN: The 21st IEEE International Symposium on Robot and Human Interactive Communication, Paris, France, 9–13 September 2012; pp. 411–417.
41. Puri, R. Gesture recognition based mouse events. *arXiv* **2014**, arXiv:1401.2058.
42. Wang, C.; Liu, Z.; Chan, S.C. Superpixel-based hand gesture recognition with kinect depth camera. *IEEE Trans. Multimed.* **2014**, *17*, 29–39. [CrossRef]
43. Garg, P.; Aggarwal, N.; Sofat, S. Vision based hand gesture recognition. *World Acad. Sci. Eng. Technol.* **2009**, *49*, 972–977.

44. Chastine, J.; Kosoris, N.; Skelton, J. A study of gesture-based first person control. In Proceedings of the CGAMES'2013 USA, Louisville, KY, USA, 30 July–1 August 2013; pp. 79–86.
45. Dominio, F.; Donadeo, M.; Marin, G.; Zanuttigh, P.; Cortelazzo, G.M. Hand gesture recognition with depth data. In Proceedings of the 4th ACM/IEEE International Workshop on Analysis and Retrieval of Tracked Events and Motion in Imagery Stream, Barcelona, Spain, 21 October 2013; pp. 9–16.
46. Xu, Y.; Wang, Q.; Bai, X.; Chen, Y.L.; Wu, X. A novel feature extracting method for dynamic gesture recognition based on support vector machine. In Proceedings of the 2014 IEEE International Conference on Information and Automation (ICIA), Hailar, China, 28–30 July 2014; pp. 437–441.
47. Jais, H.M.; Mahayuddin, Z.R.; Arshad, H. A review on gesture recognition using Kinect. In Proceedings of the 2015 International Conference on Electrical Engineering and Informatics (ICEEI), Bali, Indonesia, 10–11 August 2015; pp. 594–599.
48. Czuszynski, K.; Ruminski, J.; Wtorek, J. Pose classification in the gesture recognition using the linear optical sensor. In Proceedings of the 2017 10th International Conference on Human System Interactions (HSI), Ulsan, Korea, 17–19 July 2017; pp. 18–24.
49. Park, S.; Ryu, M.; Chang, J.Y.; Park, J. A hand posture recognition system utilizing frequency difference of infrared light. In Proceedings of the 20th ACM Symposium on Virtual Reality Software and Technology, Edinburgh, Scotland, 11–13 November 2014; pp. 65–68.
50. Jangyodsuk, P.; Conly, C.; Athitsos, V. Sign language recognition using dynamic time warping and hand shape distance based on histogram of oriented gradient features. In Proceedings of the 7th International Conference on PErvasive Technologies Related to Assistive Environments, Rhodes, Greece, 27–30 May 2014; pp. 1–6.
51. Sahoo, J.P.; Prakash, A.J.; Pławiak, P.; Samantray, S. Real-Time Hand Gesture Recognition Using Fine-Tuned Convolutional Neural Network. *Sensors* **2022**, *22*, 706. [CrossRef] [PubMed]
52. Gadekallu, T.R.; Srivastava, G.; Liyanage, M.; Iyapparaja, M.; Chowdhary, C.L.; Koppu, S.; Maddikunta, P.K.R. Hand gesture recognition based on a Harris hawks optimized convolution neural network. *Comput. Electr. Eng.* **2022**, *100*, 107836. [CrossRef]
53. Amin, M.S.; Rizvi, S.T.H. Sign Gesture Classification and Recognition Using Machine Learning. *Cybern. Syst.* **2022**. [CrossRef]
54. Kong, F.; Deng, J.; Fan, Z. Gesture recognition system based on ultrasonic FMCW and ConvLSTM model. *Measurement* **2022**, *190*, 110743. [CrossRef]
55. Saboo, S.; Singha, J.; Laskar, R.H. Dynamic hand gesture recognition using combination of two-level tracker and trajectory-guided features. *Multimed. Syst.* **2022**, *28*, 183–194. [CrossRef]
56. Alnaim, N. Hand Gesture Recognition Using Deep Learning Neural Networks. Ph.D. Thesis, Brunel University, London, UK, 2020.
57. Oudah, M.; Al-Naji, A.; Chahl, J. Computer Vision for Elderly Care Based on Hand Gestures. *Computers* **2021**, *10*, 5. [CrossRef]
58. Joseph, P. Recent Trends and Technologies in Hand Gesture Recognition. *Int. J. Adv. Res. Comput. Sci.* **2017**, *8*.
59. Zhang, Y.; Liu, B.; Liu, Z. Recognizing hand gestures with pressure-sensor-based motion sensing. *IEEE Trans. Biomed. Circuits Syst.* **2019**, *13*, 1425–1436. [CrossRef] [PubMed]
60. Mujahid, A.; Awan, M.J.; Yasin, A.; Mohammed, M.A.; Damaševičius, R.; Maskeliūnas, R.; Abdulkareem, K.H. Real-Time Hand Gesture Recognition Based on Deep Learning YOLOv3 Model. *Appl. Sci.* **2021**, *11*, 4164. [CrossRef]
61. Min, Y.; Zhang, Y.; Chai, X.; Chen, X. An efficient pointlstm for point clouds based gesture recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 5761–5770.
62. Al-Hammadi, M.; Muhammad, G.; Abdul, W.; Alsulaiman, M.; Bencherif, M.A.; Alrayes, T.S.; Mathkour, H.; Mekhtiche, M.A. Deep learning-based approach for sign language gesture recognition with efficient hand gesture representation. *IEEE Access* **2020**, *8*, 192527–192542. [CrossRef]
63. Neethu, P.; Suguna, R.; Sathish, D. An efficient method for human hand gesture detection and recognition using deep learning convolutional neural networks. *Soft Comput.* **2020**, *24*, 15239–15248. [CrossRef]
64. Asadi-Aghbolaghi, M.; Clapes, A.; Bellantonio, M.; Escalante, H.J.; Ponce-López, V.; Baró, X.; Guyon, I.; Kasaei, S.; Escalera, S. A survey on deep learning based approaches for action and gesture recognition in image sequences. In Proceedings of the 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), Washington, DC, USA, 30 May–3 June 2017; pp. 476–483.
65. Cao, C.; Zhang, Y.; Wu, Y.; Lu, H.; Cheng, J. Egocentric gesture recognition using recurrent 3d convolutional neural networks with spatiotemporal transformer modules. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 3763–3771.
66. John, V.; Boyali, A.; Mita, S.; Imanishi, M.; Sanma, N. Deep learning-based fast hand gesture recognition using representative frames. In Proceedings of the 2016 International Conference on Digital Image Computing: Techniques and Applications (DICTA), Gold Coast, QLD, Australia, 30 November–2 December 2016; pp. 1–8.
67. Zhang, X.; Li, X. Dynamic gesture recognition based on MEMP network. *Future Internet* **2019**, *11*, 91. [CrossRef]
68. Wang, S.; Song, J.; Lien, J.; Poupyrev, I.; Hilliges, O. Interacting with soli: Exploring fine-grained dynamic gesture recognition in the radio-frequency spectrum. In Proceedings of the 29th Annual Symposium on User Interface Software and Technology, Tokyo, Japan, 16–19 October 2016; pp. 851–860.

69. Funke, I.; Bodenstedt, S.; Oehme, F.; von Bechtolsheim, F.; Weitz, J.; Speidel, S. Using 3D convolutional neural networks to learn spatiotemporal features for automatic surgical gesture recognition in video. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Shenzhen, China, 13–17 October 2019; Springer: Berlin/Heidelberg, Germany, 2019; pp. 467–475.
70. Al Farid, F.; Hashim, N.; Abdullah, J. Vision Based Gesture Recognition from RGB Video Frames Using Morphological Image Processing Techniques. *Int. J. Adv. Sci. Technol.* **2019**, *28*, 321–332.
71. Al Farid, F.; Hashim, N.; Abdullah, J. Vision-based hand gesture recognition from RGB video data using SVM. In Proceedings of the International Workshop on Advanced Image Technology (IWAIT) 2019, International Society for Optics and Photonics, NTU, Singapore, 22 March 2019; Volume 11049, p. 110491E.
72. Bhuiyan, M.R.; Abdullah, D.; Hashim, D.; Farid, F.; Uddin, D.; Abdullah, N.; Samsudin, D. Crowd density estimation using deep learning for Hajj pilgrimage video analytics. *F1000Research* **2021**, *10*, 1190. [CrossRef]
73. Bhuiyan, M.R.; Abdullah, J.; Hashim, N.; Al Farid, F.; Samsudin, M.A.; Abdullah, N.; Uddin, J. Hajj pilgrimage video analytics using CNN. *Bull. Electr. Eng. Inform.* **2021**, *10*, 2598–2606. [CrossRef]
74. Zamri, M.N.H.B.; Abdullah, J.; Bhuiyan, R.; Hashim, N.; Farid, F.A.; Uddin, J.; Husen, M.N.; Abdullah, N. A Comparison of ML and DL Approaches for Crowd Analysis on the Hajj Pilgrimage. In *Proceedings of the International Visual Informatics Conference*; Springer: Berlin/Heidelberg, Germany, 2021; pp. 552–561.
75. Bari, B.S.; Islam, M.N.; Rashid, M.; Hasan, M.J.; Razman, M.A.M.; Musa, R.M.; Ab Nasir, A.F.; Majeed, A.P.A. A real-time approach of diagnosing rice leaf disease using deep learning-based faster R-CNN framework. *Peerj Comput. Sci.* **2021**, *7*, e432. [CrossRef] [PubMed]
76. Zoph, B.; Cubuk, E.D.; Ghiasi, G.; Lin, T.Y.; Shlens, J.; Le, Q.V. Learning data augmentation strategies for object detection. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 566–583.
77. Xie, Q.; Dai, Z.; Hovy, E.; Luong, M.T.; Le, Q.V. Unsupervised data augmentation for consistency training. *arXiv* **2019**, arXiv:1904.12848.
78. Islam, M.Z.; Hossain, M.S.; ul Islam, R.; Andersson, K. Static hand gesture recognition using convolutional neural network with data augmentation. In Proceedings of the 2019 Joint 8th International Conference on Informatics, Electronics & Vision (ICIEV) and 2019 3rd International Conference on Imaging, Vision & Pattern Recognition (icIVPR), Spokane, WA, USA, 30 May–2 June 2019, pp. 324–329.
79. Mungra, D.; Agrawal, A.; Sharma, P.; Tanwar, S.; Obaidat, M.S. PRATIT: A CNN-based emotion recognition system using histogram equalization and data augmentation. *Multimed. Tools Appl.* **2020**, *79*, 2285–2307. [CrossRef]
80. Rashid, M.; Bari, B.S.; Yusup, Y.; Kamaruddin, M.A.; Khan, N. A Comprehensive Review of Crop Yield Prediction Using Machine Learning Approaches With Special Emphasis on Palm Oil Yield Prediction. *IEEE Access* **2021**, *9*, 63406–63439. [CrossRef]
81. Rashid, M.; Sulaiman, N.; PP Abdul Majeed, A.; Musa, R.M.; Bari, B.S.; Khatun, S. Current status, challenges, and possible solutions of EEG-based brain-computer interface: A comprehensive review. *Front. Neurorobotics* **2020**, *14*, 25. [CrossRef]
82. Mathew, A.; Amudha, P.; Sivakumari, S. Deep Learning Techniques: An Overview. In Proceedings of the International Conference on Advanced Machine Learning Technologies and Applications, Manipal, India, 13–15 February 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 599–608.
83. Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; MIT Press: Cambridge, MA, USA, 2016.
84. Ji, S.; Xu, W.; Yang, M.; Yu, K. 3D convolutional neural networks for human action recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *35*, 221–231. [CrossRef]
85. Liu, Z.; Zhang, C.; Tian, Y. 3D-based deep convolutional neural network for action recognition with depth sequences. *Image Vis. Comput.* **2016**, *55*, 93–100. [CrossRef]
86. Sun, L.; Jia, K.; Yeung, D.Y.; Shi, B.E. Human action recognition using factorized spatio-temporal convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 4597–4605.
87. Escorcia, V.; Heilbron, F.C.; Niebles, J.C.; Ghanem, B. Daps: Deep action proposals for action understanding. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; Springer: Berlin/Heidelberg, Germany, 2016; pp. 768–784.
88. Mansimov, E.; Srivastava, N.; Salakhutdinov, R. Initialization strategies of spatio-temporal convolutional neural networks. *arXiv* **2015**, arXiv:1503.07274.
89. Baccouche, M.; Mamalet, F.; Wolf, C.; Garcia, C.; Baskurt, A. Sequential deep learning for human action recognition. In Proceedings of the International Workshop on Human Behavior Understanding, Amsterdam, The Netherlands, 16 November 2011; Springer: Berlin/Heidelberg, Germany, 2011; pp. 29–39.
90. Feichtenhofer, C.; Pinz, A.; Zisserman, A. Convolutional two-stream network fusion for video action recognition. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1933–1941.
91. Shou, Z.; Wang, D.; Chang, S.F. Temporal action localization in untrimmed videos via multi-stage cnns. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1049–1058.
92. Varol, G.; Laptev, I.; Schmid, C. Long-term temporal convolutions for action recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 1510–1517. [CrossRef]

93. Neverova, N.; Wolf, C.; Taylor, G.W.; Nebout, F. Multi-scale deep learning for gesture detection and localization. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; Springer: Berlin/Heidelberg, Germany, 2014; pp. 474–490.
94. Wang, L.; Qiao, Y.; Tang, X. Action recognition with trajectory-pooled deep-convolutional descriptors. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 4305–4314.
95. Han, S.; Mao, H.; Dally, W.J. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. *arXiv* **2015**, arXiv:1510.00149.
96. Zhang, B.; Wang, L.; Wang, Z.; Qiao, Y.; Wang, H. Real-time action recognition with enhanced motion vector CNNs. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2718–2726.
97. Xu, Z.; Zhu, L.; Yang, Y.; Hauptmann, A.G. Uts-cmu at thumos 2015. *Thumos Chall.* **2015**, *2015*, 2.
98. Gkioxari, G.; Malik, J. Finding action tubes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 759–768.
99. Escalante, H.J.; Morales, E.F.; Sucar, L.E. A naive bayes baseline for early gesture recognition. *Pattern Recognit. Lett.* **2016**, *73*, 91–99. [CrossRef]
100. Xu, X.; Hospedales, T.M.; Gong, S. Multi-task zero-shot action recognition with prioritised data augmentation. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; Springer: Berlin/Heidelberg, Germany, 2016; pp. 343–359.
101. Montes, A.; Salvador, A.; Pascual, S.; Giro-i Nieto, X. Temporal activity detection in untrimmed videos with recurrent neural networks. *arXiv* **2016**, arXiv:1608.08128.
102. Nasrollahi, K.; Escalera, S.; Rasti, P.; Anbarjafari, G.; Baro, X.; Escalante, H.J.; Moeslund, T.B. Deep learning based super-resolution for improved action recognition. In Proceedings of the 2015 International Conference on Image Processing Theory, Tools and Applications (IPTA), Orleans, France, 10–13 November 2015; pp. 67–72.