Neural Computing and Applications (2023) 35:8143–8156
https://doi.org/10.1007/s00521-022-08090-8

ORIGINAL ARTICLE

Real-time automated detection of older adults' hand gestures in home and clinical settings

Guan Huang[1] · Son N. Tran[1] · Quan Bai[1] · Jane Alty[2,3]

Received: 21 March 2022 / Accepted: 22 November 2022 / Published online: 10 December 2022

Abstract

There is an urgent need, accelerated by the COVID-19 pandemic, for methods that allow clinicians and neuroscientists to remotely evaluate hand movements. This would help detect and monitor degenerative brain disorders that are particularly prevalent in older adults. With the wide accessibility of computer cameras, a vision-based real-time hand gesture detection method would facilitate online assessments in home and clinical settings. However, motion blur is one of the most challenging problems in the fast-moving hands data collection. The objective of this study was to develop a computer vision-based method that accurately detects older adults' hand gestures using video data collected in real-life settings. We invited adults over 50 years old to complete validated hand movement tests (fast finger tapping and hand opening–closing) at home or in clinic. Data were collected without researcher supervision via a website programme using standard laptop and desktop cameras. We processed and labelled images, split the data into training, validation and testing, respectively, and then analysed how well different network structures detected hand gestures. We recruited 1,900 adults (age range 50–90 years) as part of the TAS Test project and developed UTAS7k—a new dataset of 7071 hand gesture images, split 4:1 into clear: motion-blurred images. Our new network, RGRNet, achieved 0.782 mean average precision (mAP) on clear images, outperforming the state-of-the-art network structure (YOLOV5-P6, mAP 0.776), and mAP 0.771 on blurred images. A new robust real-time automated network that detects static gestures from a single camera, RGRNet, and a new database comprising the largest range of individual hands, UTAS7k, both show strong potential for medical and research applications.

Keywords TAS Test · Hand gesture classification · Similar object detection · Motion blur · Dementia · Bradykinesia

1 Introduction

In recent years, hand gesture detection has been increasingly explored in human computer interaction research, including sign language detection, video games and virtual reality. The rapid development of deep learning [1, 2] has significantly improved the accuracy of hand gesture detection; for example, researchers have used 3DCNN models to accurately classify hand gestures used in Arabic sign language with 90% accuracy [3].

There is now growing interest in how these technologies can be applied to medical and neuroscience research applications as there is an urgent need, accelerated by the COVID-19 pandemic, for methods that allow remote evaluation of hand movements. Hand movement assessments play a key part in the detection and monitoring of brain disorders such as Parkinson's and stroke. These disorders are particularly prevalent in older adults and usually

✉ Son N. Tran
sn.tran@utas.edu.au

Guan Huang
guanh@utas.edu.au

Quan Bai
quan.bai@utas.edu.au

Jane Alty
jane.alty@utas.edu.au

[1] College of Sciences and Engineering, University of Tasmania, Sandy Bay, TAS 7005, Australia

[2] Wicking Dementia Research and Education Centre, University of Tasmania, Hobart, TAS 7000, Australia

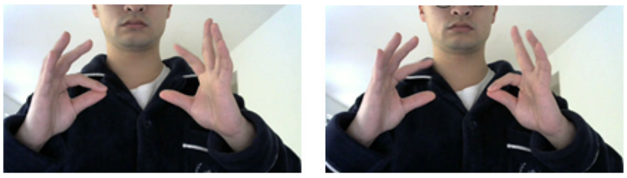
[3] School of Medicine, University of Tasmania, Hobart, TAS 7000, Australia

require participants to attend clinics or research institutes for face-to-face tests. This is problematic for people who live in rural or remote locations, those with limited mobility and for the majority of patients and research participants during the COVID-19 pandemic. With the wide accessibility of computer cameras, a vision-based method that can detect hand gestures in real time would facilitate online assessments in the home and clinical settings, and this would transform the accessibility and efficiency of medical assessments and research studies.

Apraxia is a neurological disorder characterised by difficulties carrying out precise movements despite the physical ability to perform them. It is usually due to impaired brain connections that integrate the planning, sequencing and motor–sensory integration of movement. Causes in adults include stroke and neurodegenerative disorders, such as Alzheimer's disease and Parkinson's disease [4]. We are developing a new online test, the TAS Test, that analyses hand movement features [5] associated with ageing and neurodegenerative disorders, in a large established cohort of older adults. Participants access the online test using their own laptop or desktop computer via a website and then follow a series of instructions to record their hand movements with a webcam. The test is designed to be completed remotely from the research centre without any researcher supervision [6]. A robust real-time automated method is important for clinician and researchers to monitor whether the participants are following the instructions and therefore to further evaluate and analyse the level of apraxia in the hand movement.

However, real-time gesture recognition of fast-moving hands is challenging for several reasons. First, the validated hand movements tests, such as finger tapping and whole hand opening–closing require participants to repeat these movements 'as fast as possible' and this creates motion blur, especially for home computer cameras that tend to have a relatively low rate of frames per second (fps). Second, accurately recognising similar gestures such as finger and thumb together (closed position) at the start of the finger tapping cycle or a few centimetres apart (open position) partway through the finger tapping cycle results in inherent errors and confusion, which decreases the detection accuracy. Third, in the home and clinic settings, the backgrounds are typically cluttered and there are variations in ambient lighting and distance of the hands to the camera.

The hand movement tests include repetitive finger tapping (tapping the index finger against the thumb, in the phase and anti-phase) and repetitive hand opening–closing (of all the digits). Both are well-validated tests for evaluating human movement function. Figure 1 illustrates how the anti-phase (or alternate) finger tapping test is performed; the participants are instructed to switch between 'Gesture 1' and 'Gesture 2' quickly and repeatedly [7].



(a) Gesture 1 **(b)** Gesture 2

Fig. 1 Two images were taken from a video recording of the anti-phase (alternate) finger tapping test. **a** Gesture 1 with the right-hand finger and thumb opposed and the left-hand finger and thumb separated. **b** Gesture 2 with the left-hand finger and thumb opposed and the right-hand finger and thumb separated

The overall aim of this study was to develop a robust method to detect fast-moving hand gestures in real time. Our first objective was to develop a large dataset of hand gestures collected in home environments and clinical settings (with real-life cluttered backgrounds), and split into clear and motion-blurred images. Our second objective was to develop an accurate method for discriminating similar hand gestures in clear and blurred images while remaining real time and compare the accuracy of this to other established methods.

In this study, we establish a new dataset, UTAS7k, with 20 per cent of blurred images included. We compare the detection accuracy of different network structures on hand gesture classification, where we find multi-scale detection technique was effective. To develop an optimal network structure for hand gesture detection, we embedded different types of neural networks into the detector network. We implemented an attention-based hand gesture network to detect the hand gestures performed in the hand movement tests. We developed a new model, RGRNet, for fast-moving hand gesture detection, inspired by CSPDarknet-53 [8]. We have implemented the multi-scale detection technique and embedded one more detection head for the detection of hands in different sizes. We also adopted attention layers in the feature extraction blocks to increase the prediction capability. To further increase the performance of detection, data augmentation was employed during training, including mosaic and left or right flip.

Our experiments revealed that multi-scale detection techniques help to increase the overall performance of similar gesture classification. Also, the experimental results show that our new model, RGRNet improved the accuracy of similar gesture classification on clear images and performs more robustly on both clear and motion-blurred hand gesture images.

The key contributions of this research study are as follows:

1. We develop a novel model, RGRNet, with attention and transformer layers to improve the classification performance on similar gestures and blurred images extracted from hand movement videos. Our model achieved near real-time hand gesture detection and classification with a total processing speed of 18.8ms.
2. We establish UTAS7k, a real-world dataset comprising more than 7,000 images of similar hand gestures with cluttered backgrounds and split 4:1 into clear and motion-blurred images. Then, we provide a comprehensive comparison of the classification accuracy of different network structures on the hand gesture videos with motion blur and similar gestures and show that our model, RGRNet, has a better performance than the state-of-the-art network structure. All of these contributions have strong potential for medical, neuroscience and computer science research applications.

The organisation of the paper is as follows: in Sect. 2, we summarise the literature related to our research, including hand gesture detection methods, object detection algorithms and network structures for object detection. In Sect. 3, we describe our network structure in detail, including the structure of each block. In Sect. 4, we describe the collection and processing steps to develop UTAS7k, our dataset of hand gestures. In Sect. 5, we present the experiment methods and results of classification accuracy and detection speed of our new network, RGRNet, compared to a range of other popular network structures. Finally, in Sect. 6, we summarise the findings from our work and discuss future directions.

2 Related work

2.1 Hand gesture detection methods

Methods to detect hand gestures are generally categorised into two types: wearable device-based gesture recognition or computer and vision-based gesture recognition. The first method typically measures the angle between fingers and proximal interphalangeal joints to estimate the gestures [9] although the activity of the muscles in the digits and upper limb has also been used [10]. A range of wearable devices have been employed, including data gloves (with embedded sensors), tactile switches, optical goniometers and resistance sensors to measure the bending of finger joints [11]. This approach mainly focuses on increasing the accuracy to pinpoint the position of the hand in the 3D model. However, the main limitations of the wearable sensor approach are accessibility, cost and infection. The clinician or researcher needs to find a robust method of delivering the sensors to patients or participants with clear instructions, or bring the participants into the research laboratory setting. Also, sensors can be very expensive; for example, commercial wearable data gloves typically cost in the range of $1000 to $20,000 each [12]. Furthermore, any multi-user wearable devices will have infection control issues as there is a need for thorough cleaning between participants. All of these barriers limit the usefulness of wearable devices for hand gesture detection in medical and large scale studies.

The second method, the vision-based method, holds much more potential for remote or large scale studies as participants' hand gestures are captured as image data by video cameras and then processed and analysed through computer vision algorithms [13]. Before the popularity of CNNs in hand gesture recognition, the traditional vision-based approach focuses on extracting image features and then using a classifier to differentiate features into different gestures. Statistical methods were the most widely used. For example, Lee and Kim [14] first introduced Hidden Markove Model (HMM) to calculate likelihood of the detected gestures. Many subsequent efforts have been made to improve the classification performance, for example, IOHMM [15] and the combination of HMM and recurrent neural networks (RNN) [16]. Some work focus on how to extract features effectively, such as stochastic context-free grammar (SCFG) [17].

With the development of deep learning and convolutional neural network (CNN), researchers have been employing CNNs for hand gesture recognition thanks to their ability to learn visual features from images; hence, feature extraction is not required [18]. Real-time hand gesture recognition has benefited greatly from this as many popular object detection and image classification algorithms have been developed recently. They include inception V2 for MITI hand dataset [19], SSD for American Manual Alphabet detection [20], YOLOV3 on a custom hand gesture dataset [21] and Temporal Segment Networks (TSN) [22] for IPN Hand dataset [23]. Many of those approaches have achieved high accuracy, confirming that vision-based hand gesture detection methods can be a reliable method for hand gesture detection problems. Unlike static hand gesture detection, dynamic hand gesture recognition is more challenging because blurriness boundaries of the hand gestures [24]. Deep learning-based methods also show promising results on dynamic gesture recognition, Kopuklu et al. implemented ResNet-10 and achieved 77.39% accuracy on nvGesture Dataset [25]. Do et al. also achieved 96.07% accuracy on a custom dynamic hand gesture dataset by using a ConvLSTM model [26]. However, previous literature shows limited information about similar gesture detection and how real-world hand gesture classification problem were analysed.

Object detection algorithms used in recent years have tended to be either two-stage detectors or one-stage detectors and each has its own advantage: two-stage detectors are good at improving the accuracy of detection and one-stage detectors generally have faster detection speeds. YOLO, which was first introduced by Redmon in 2016 [27], is regarded as one of the most successful one-stage object detectors and has been widely adopted for real-time hand gesture detection. For example, Ni et al. implemented a hand gesture detection system based on YOLOV2 for hand gesture recognition in scenes with a cluttered background [28] and Mujahid et al. proposed a YOLOV3 system for real-time gesture recognition in detecting hand gestures and then denoted numbers from 1 to 5 [21].

The YOLO object detection model has been updated and improved constantly. From the first version through to the latest version, YOLOV5, many techniques such as batch normalisation, anchor boxes, multi-scale training, feature pyramid networks for object detection, mosaic data augmentation and model size reduction have been implemented to improve the performance [8, 29–31]. Nevertheless, despite the successes of YOLO, some researchers have found it lacks capabilities in detecting small objects such as fingers of the hand or small images of pedestrians on the road [32].

In summary, the development of vision-based object detection techniques has dramatically improved both the accuracy and speed of hand gesture detection, but there remains a lack of research into real-world challenges. Two key challenges include how to detect hand gestures in real time when there is motion blur and how to discriminate very similar gestures. These challenges are commonplace in medical and neuroscience applications, especially during the COVID-19 pandemic, when patients and participants are increasingly using their own laptop cameras, or clinic webcams to collect hand data remotely.

2.2 Feature extraction network

So far, the detection of hand gestures in real time has relied heavily on feature extraction networks as the backbone for the majority of solutions. Such networks are commonly used to extract deep features from the images. For example, ResNet [33] has been widely adopted as a feature extractor and backbone in many one-stage detectors, including RetinaNet [34] and SSD [35]. ResNet introduced a shortcut connection that guaranteed the gradient would not be vanished. EfficientNet [36] is another popular feature extraction network and this network used a neural architecture search method to explore the optimistic model depth, width and resolution of input images. This provides a way to adaptively scale the model to optimise the computational cost for different computer vision tasks.

In order to extract more useful features, recent approaches have employed transformers to pay attention to the discriminable information in the inputs. Attention was originally designed as a useful tool in natural language processing (NLP), but has shown potential in wider applications. For computer image classification, it enables the neural network to learn the relevant information of the images for the tasks and increase the performance [37].

3 Rapid gesture recognition net (RGRNet)

3.1 Network structure

We proposed a novel network structure called 'Deep Robust Hand Gesture Network' (RGRNet), which includes attention mechanisms and multi-scale detection techniques to increase the accuracy for detecting fast-moving hands and for classifying similar gestures. Our proposed network consists of three blocks as shown in Fig. 2: Block 1 is designed as the feature extractor, Block 2 is for feature fusion and Block 3 is for detection.

Block 1 - Feature extractor

Our network structure was designed as a classic one-stage detector framework. Our feature extractor includes traditional CSPNet structure [38] and attention blocks. In Block 1, the size of the convolution kernel in front of each CSP module is 3x3 with a stride equal to 2. This architecture allows the network to have different sizes of feature maps from top to the bottom.

Block 2 - Feature fusion

In Block 2, feature pyramid networks (FPN) [39] and PAnet (PAN) [40] structure were employed. The FPN can effectively propagate semantic visual information in a bottom-up manner while PAN would enhance the localisation of discriminable features in the lower layer and link them to the top layer. The idea to combine the two different structures would encourage better parameter aggregation between different layers and a more effective fusion of visual features.

Block 3 - Detection

Block 3 inherits the output from the feature fusion block, followed by two layers of convolution. The final outputs of Block 3 will include: (1) bounding boxes and their confidence scores and (2) a SoftMax layer of N, where N is the number of gestures.

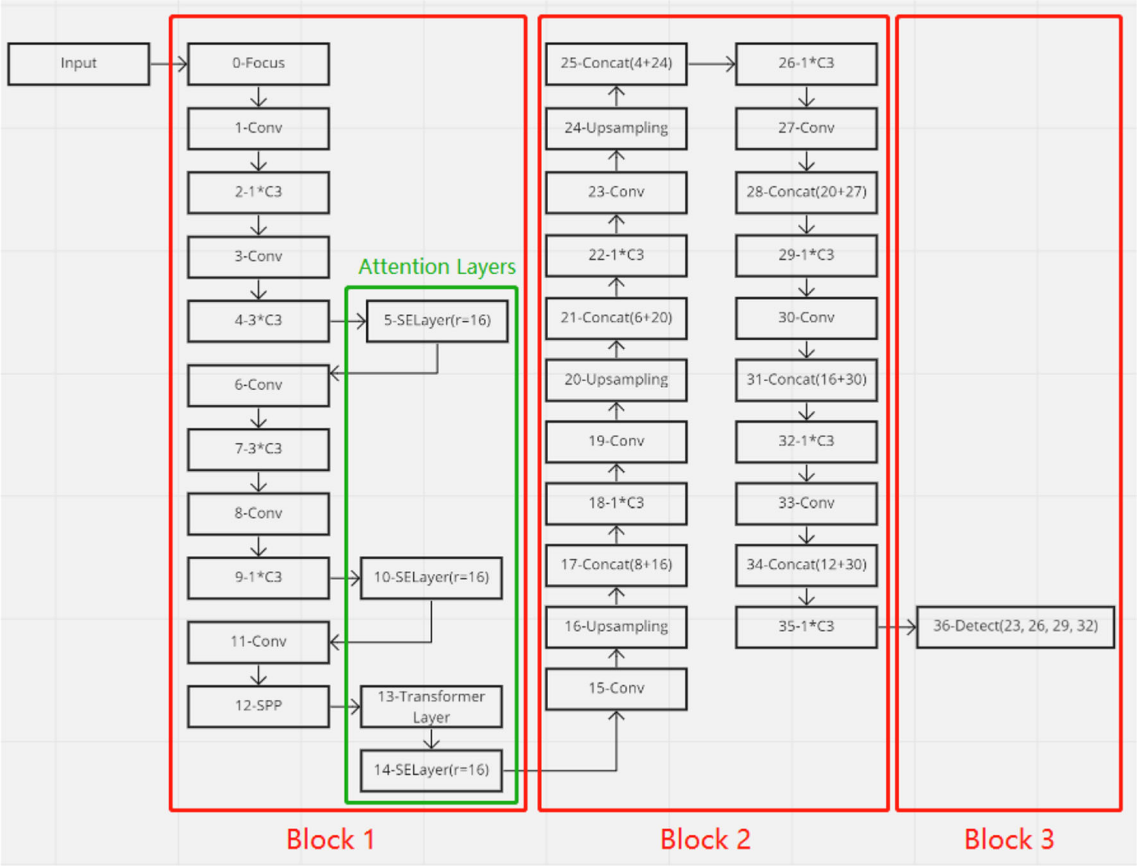


Fig. 2 Rapid gesture recognition net (RGRNet)

Focus layer

The focus layer [31] is the first layer in the feature extraction network. A 608x608 image after the focus layer will be sliced into 4 different parts; then, we use 32 convolution kernels to convert them into $304\times304\times32$ feature maps. This technique was also used by Tal et al. [41] who called it 'SpaceToDepth', and stated it allowed the network to rearrange spatial data into depth. Such a component will thus help our network to reduce the resolution of the input images and save computational costs.

Conv layer

The Conv layer is a fundamental building block in our network. In the Conv layer, the input feature maps will go through a 3x3 convolution with strike equals 2, followed by batch normalisation and a SiLU activation function [42].

C3 layer

The C3 layer is a stackable layer in our network, which consists of CBL and multiple CSP units. In this layer, CBL refers to convolution, batch normalisation and Leaky ReLU activation function. In the CSP unit, the input feature map will go through two CBL blocks and then add the previous output to its original feature map becoming the output. This approach ensures that the amount of information under the feature maps of the image is increased while guaranteeing that the dimensions of the feature maps are not increased. This operation will increase the amount of information and is beneficial to the final classification of the image.

SPP layer

In the SPP layer [8], we use four different types of max-pooling to fuse the feature maps. They are 1×1, 5×5, 9×9 and 13×13 max-pooling. This block enables the information from different sizes of feature maps to be combined.

Upsampling layer

The upsampling layer enables the feature map size to be increased. In our network, we adopted the nearest neighbour in the upsampling calculation.

Concat layer

The concat layer refers to an operation that combines the feature maps in different sizes. This layer enable us combine features from different layers and fuse to a new feature.

3.2 Attention layers

The core element in our architecture is the attention module, as shown in the green box of Fig. 2. This module consists of several attention layers, including squeeze-and-excitation (SE) layer and transformer layer.

Squeeze-and-excitation (SE) layer

The squeeze-and-excitation (SE) layer enhances the ability of models to learn correlations between visual channels (R–G–B, H–S–V, etc.). The SE-Block was first proposed in SENet [43] and showed better performance than ResNet. Although the SE layer will slightly increase the computation costs, the performance degradation is within acceptable limits, and the loss of SENet is not significant in terms of GFLOPs, the number of parameters and run-time experiments. The architecture of the SE layer is demonstrated in Fig. 3a. The SE layer is normally embedded after a traditional convolution block. Firstly, we use global average pooling to reduce the dimensionality of the feature maps from 3D to 1D, this step can also be referred to as 'squeeze'.

The squeeze operation is calculated as the following equation :

$$Z_c = F_{sq}(U_c) = \frac{1}{H \times W}\sum_{i=1}^{H}\sum_{j=1}^{W} U_c(i,j) \tag{1}$$

where F_{sq} refers to the feature map after 'squeeze', U_c is the transformation output from the previous layer, H and W are height and width, respectively. By using the average pooling method, all the information contained in this feature map is averaged. 'Excitation' is done by two fully connected layers. The first fully connected layer will squeeze the number of channels from C to C/r, where r refers to the reduction ratio. The second fully connected layer is adopted to ensure the feature map can be returned to its original channel size. This attention mechanism allows the network to focus more on the most informative channel features and suppress the less important ones.

Transformer layer

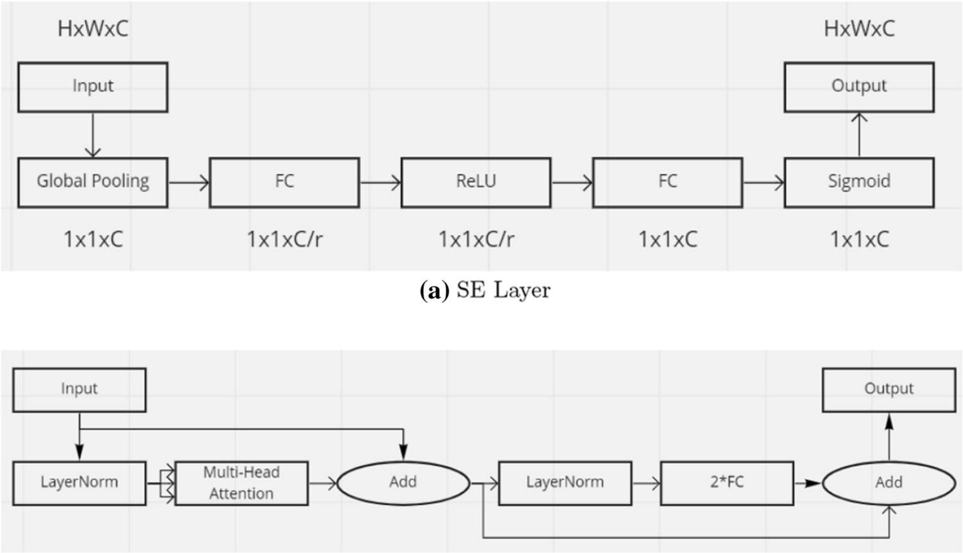
The transformer layer is shown in Fig. 3b, which is inspired by the vision transformer. The vision transformer has implemented a skip-connection-based network block and demonstrated having better performance than the state of the art on image classification tasks [37].

In our model, the transformer consists of multilayer perception, two normalisation layers and one multi-head attention. It enables the neural network to extract global information from image features. By adding attention, the global relationship and distinctions between five hand gestures frequently used in the finger tapping and hand opening tasks can be learned from an input image. In the transformer layer, the global information is learnt through the similarity calculation. Vaswani et al. have used Query and Key-Value pairs to represent the similarity of the features [44], where the similarity is calculated as:

$$f(Q,k_i), i = 1,2,\ldots,m \tag{2}$$

The similarity function f is normalised by applying a SoftMax operator, followed by calculating the weighted sum of all the values in V to obtain the attention vector. We can use the following equation to calculate attention in the transformer:

Fig. 3 SE layer and transformer layer



(a) SE Layer

(b) Transformer Layer

$$Attention(Q,K,V) = softmax\left(\frac{Q^TK}{\sqrt{dk}}\right)V, \tag{3}$$

where $\frac{1}{\sqrt{dk}}$ is a scaling factor. With multi-head, we are able to perform attention multiple times in parallel (we set 4 heads in our experiment) without sharing parameters. The fusion of the attention and transformer layers enables our network to focus on specific areas for solving particular tasks (hand gesture recognition), rather than evaluating the whole image.

4 Dataset

4.1 UTAS7k dataset—a new dataset for hand gesture detection

Subjects and setting

As part of the TAS Test [5] project, we invited adults aged 50 years and older from an established cohort in Tasmania, Australia (The ISLAND Project) [45] to perform a series of hand movement tests. The TAS Test project aims to track movement and cognitive changes associated with ageing and degenerative brain disorders over a 10-year period. Ethical approval has been granted for the TAS Test study by the Human Research Ethics committee at the University of Tasmania (reference H0021660), and all participants provide informed online consent.

Hand movement data were collected via TAS Test, an online programme, that uses a short demonstration video to instruct participants how to perform each of the hand movement tests and then records their hand movements as they complete the test using a standard laptop camera or desktop webcam (typically with 30 fps). The test is designed to be completed without any in-person researcher assistance or supervision.

So far, 1,900 participants aged between 50 and 90 years have completed a range of hand movement tests, with the majority completing tests in their own home and some in the clinical research facility at the University of Tasmania.

Dataset processing

The hand movements video consisted of a sequence of images (video frames). By stopping at a specific frame in the sequence of a hand movement video frame, we could extract a still image. Our model works by detecting a single frame in the video and then returning the result of the detection to each frame in the video. In total, more than 20,000 image frames of hand gestures were collected. In this dataset, we processed and labelled more than 7,071 images for training, validation and testing. Most data frames were collected through high-definition video (720P) or full high-definition (1080P) web cameras. Their resolutions are 1280 x 720 pixels and 1920 x 1080 pixels, respectively. To unify the size of the input image for our network and accelerate the training speed, we scaled down the image systematically (via FFmpeg) and scaling down time is not calculated; however, the average processing time for 720P video is 7.8ms/frame and 1080P at 9.1ms/frame. The data were split into 4:1 with 5996 clear images and 1075 blurred images and this dataset was named 'UTAS7k'. Table 1 outlines how the UTAS7k dataset compares to other established hand datasets, and highlights that it has a far larger population size of individual hands (n = 1900 participants) than previous datasets (n < 643 participants). Moreover, we are the only group to have also included data with motion blur.

4.2 Developing the UTAS7k dataset

4.2.1 Hand gestures

To establish the UTAS7k dataset, 5 different hand gestures were extracted from the fast finger tapping and hand opening–closing hand movement tests and called these 'open', 'close', 'pinch open', 'pinch close' and 'flip' as shown in Fig. 4.

4.2.2 Motion blur

Quantifying the blurriness of the images is essential for us to classify and pre-process the training data. Pech-Pacheco et al. proposed a method to calculate the blurriness of the images by calculating the standard deviation of a convolution operation after a Laplace mask [51]. The Laplace mask equation is listed below:

$$LAP(I) = \sum_{n}^{M}\sum_{m}^{N}|L(x,y)|, \tag{4}$$

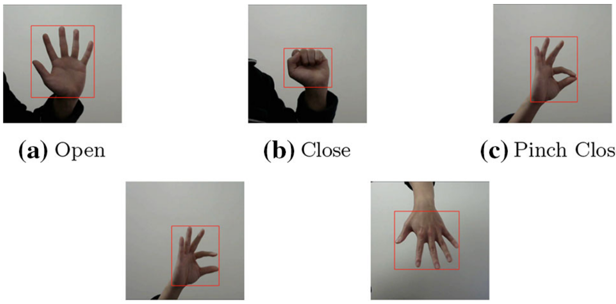
where L(m, n) is the convolution of the input image I(m,n) with the mask L and the mask is calculated by the following equation:

$$L = \frac{1}{6}\begin{pmatrix} 0 & -1 & 0 \\ -1 & 4 & -1 \\ 0 & -1 & 0 \end{pmatrix} \tag{5}$$

Figure 5 shows the blurriness score after calculation; if an image has a high variance (low blurriness), it means there are many edges in the image as is commonly seen in a normal, accurately focused picture. On the other hand, if the image has a small variance (high blurriness), then there are fewer edges in the image, which is typical of a motion-blurred image.

Table 1 Table of datasets for hand gesture detection comparison

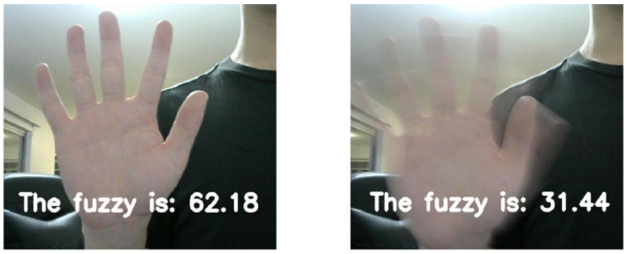
Datasets	Total images	Population	Hand side	Left–right	Age recorded	Include motion blur
11k Hands [46]	11,076	190	Palm-dorsal	Both	Yes	No
CASIA [47]	5,502	312	Palm	Both	No	No
Bosphorus [48]	4,846	642	Palm	Both	No	No
lITD [49]	2,601	230	Palm	Both	No	No
GPDS150hand [50]	1,500	150	Palm	Right	No	No
UTAS7k (ours)	7,071	1,900	Palm-dorsal	Both	Yes	Yes



(a) Open **(b)** Close **(c)** Pinch Close

(d) Pinch Open **(e)** Flip

Fig. 4 Five hand gestures in the UTAS7k dataset



(a) Clear image **(b)** Motion blur

Fig. 5 Image blurriness calculation. The ‘fuzzy’ score is a number that quantifies the average quality of the image with Fuzzy $\geq$ 50 indicating the image is clear, Fuzzy score $<$ 50 indicating motion blur

Table 2 Number of images included in the sub-datasets of UTAS7k - split into the clear dataset, motion-blurred dataset and testing dataset

Dataset	Total	Clear	Motion blur
Clear dataset	5,376	5,376	0
Motion-blurred dataset	6,450	5,376	859
Testing dataset	621	497	124

The variance (blurriness) of 7,071 images was calculated and classified into two categories: they are clear (Fuzzy $\geq$ 50) and blurred images (Fuzzy $<$ 50), respectively. Table 2 shows the number of clear images and blurred images in each dataset.

5 Experiments

5.1 Setup

Presetting In our experiment, the performance of the models was compared with state-of-the-art network structures in object detection. The following competitors were tested: DarkNet-53 [30], GhostNet [52], TinyNet [30], CSPDarknet-53 [31], MobileNetv3-small[53], EfficientNet-B1 [36] and RGRNet. (Network structure is shown in Fig. 2.) For fair comparison, all models are implemented using YOLOv5 framework. We then conduct an intensive evaluation of our model using multiple metrics including the number of parameters, as what will be discussed in the next section.

Data preparation 80 per cent of the data were split for training and 20 per cent for validation. The testing dataset was an independent dataset comprising 621 images (of 5 different gestures, 80% clear and 20% motion blurred) that were not seen by the models before. In the default settings, we set the training images as image size 640 × 640 pixels.

Training techniques To improve the quality of the training of all models, several popular data augmentation techniques were implemented, including translate, scale, flip from left to right and mosaic augmentation. This process would enrich the data which is needed for deep learning.

Stochastic gradient descent (SGD) was employed as the optimising function with a decaying learning rate of 0.0005 where the initial learning rate is set as 0.01. Before that, the training process started with very low learning rate for warm-up training to help the models gradually adapt to the data.

5.2 Evaluation metrics

mAP To evaluate whether the detection was successful, 'IOU' was used to describe the intersection over union (IOU) area between ground truth and predicted bounding boxes. IOU indicates how accurate the bounding boxes are in terms of localising objects. Normally, a threshold t (in our experiments, we set $t = 0.65$) will be assigned to determine whether the detection is successful, i.e. if the IOU of a detected bounding box is larger than t then it will be accepted as a true bounding box; otherwise, it will be classified as incorrect. mAP@0.5:0.95 was adopted as our primary evaluation metric for detection accuracy in our experiments. mAP@0.5:0.95 is the primary evaluation metric from the MS COCO challenge [54] and denotes the mAP at different thresholds (from 0.5 to 0.95 in steps of 0.05), which is calculated by the following equation:

$$mAP@0.5:0.95 = (mAP@0.5 + mAP@0.55 + \ldots + mAP@0.95)/10 \tag{6}$$

GigaFLOPS (GFLOPs) We employed Giga floating point operations per second (GFLOPs) to evaluate the computational cost. Generally, the more GFLOPs a model has, the greater the cost of the computer to run the model.

Parameters The number of parameters often determines the learning capacity of a model, the more parameters a model has, the more learning capacity it poses. The unit for this evaluation metric is 'M', meaning a million parameters.

Storage Size The storage size evaluates the amount of space for the model to be stored in the computer. The unit for storage size is Megabyte (MB).

Inference speed per image Inference speed per image evaluate the inference time for the model to process an image with a 640 × 640 pixel image size. We used milliseconds (ms) as the unit.

5.3 Experimental results

5.3.1 Performance analysis on datasets comprising clear images of hand gestures

The network comparisons are displayed in Table 3, where we evaluate different network structures performed on the testing dataset and the fastest inference speed and highest accuracy are highlighted. Our network RGRNet had a mean average precision of 0.782 which was superior to all other network structures on the clear images of hand gesture dataset. The end-to-end image processing speed at 720P image size is 17.5 ms (57FPS) and 18.8 ms (57FPS) at 1080P image size, which is still within the range of near real-time detection ($\geq$30FPS). Although the processing speed was longer at 9.7ms per image, as a one-stage detector, the inference time is already within the range of real-time detection. Adding a multi-scale detection and attention layer will increase the parameters of the network considerably. EfficientNet is one of the SOTA convolutional neural networks by setting certain parameter values to balance the depth, width and input image size of the convolutional neural network. We applied efficientNet-B1 to our data and found that it can achieve a decent result, 0.757 mAP and GFLOP (6.7) with only 9.98 M parameters. CSPNet was designed to minimise duplicate gradient information within the network and reduce the complexity of the network. In our experiment, CSPDarknet-53 also shows effectiveness on hand gesture classification by achieving 0.753 mAP and 7.5 ms image inference speed at an image size 640×640. We have taken the extra step of adding in a transformer layer and an SE layer and achieved 0.782 mAP and 9.7 ms inference speed, which is significantly higher than our baseline models, MobileNetv3-Small (0.701), EfficientNet-B1 (0.757) and CSPNet-53 (0.753).

According to the experiment results, we categorised network structures into three types: medium structures, large structures and light structures. Large network structures, such as CSPDarknet-53, EfficientNet-B1 and Darknet-53, had similar detection accuracy on our dataset and

Table 3 Network structure comparison on testing dataset (image tested on NVIDIA GTX 1660 SUPER)

Network structure	Parameters	GFLOPs	Storage size	Inference speed	mAP@0.5:0.95
GhostNet	4.17 M	9.2	7.8 MB	7.4 ms	0.735
CSPDarknet-53	7.27 M	16.9	14.00 MB	7.5 ms	0.753
MobileNetv3-Small	3.55 M	6.3	4.21 MB	**3.0 ms**	0.701
TinyNet	2.18 M	3.3	5.85 MB	3.1 ms	0.703
Darknet-53	6.99 M	19.0	14.20 MB	7.4 ms	0.755
EfficientNet-B1	9.98 M	6.7	19.40 MB	9.5 ms	0.757
YOLOV5-P6	12.36 M	16.7	25.1 MB	9.5 ms	0.776
RGRNet (ours)	12.54 M	17.0	25.5 MB	9.7 ms	**0.782**

Table 4 Network performance with noisy dataset (image tested on GTX 1660 SUPER)

Network structure	mAP without noise	mAP with noise	mAP dropped	Process speed per image
CSPDarknet(YOLOV5)	0.753	0.745	0.008	7.4ms
GhostNet	0.735	0.726	0.009	7.5ms
MobileNetv3	0.701	0.694	0.007	**3.0ms**
Darknet-53	0.755	0.747	0.008	7.4ms
EfficientNet-B1	0.757	0.745	0.012	9.5ms
TinyNet	0.703	0.701	**0.002**	3.1ms
CSPDarknet-P6	0.776	0.769	0.007	9.5ms
RGRNet (ours)	**0.782**	**0.771**	0.011	9.7ms

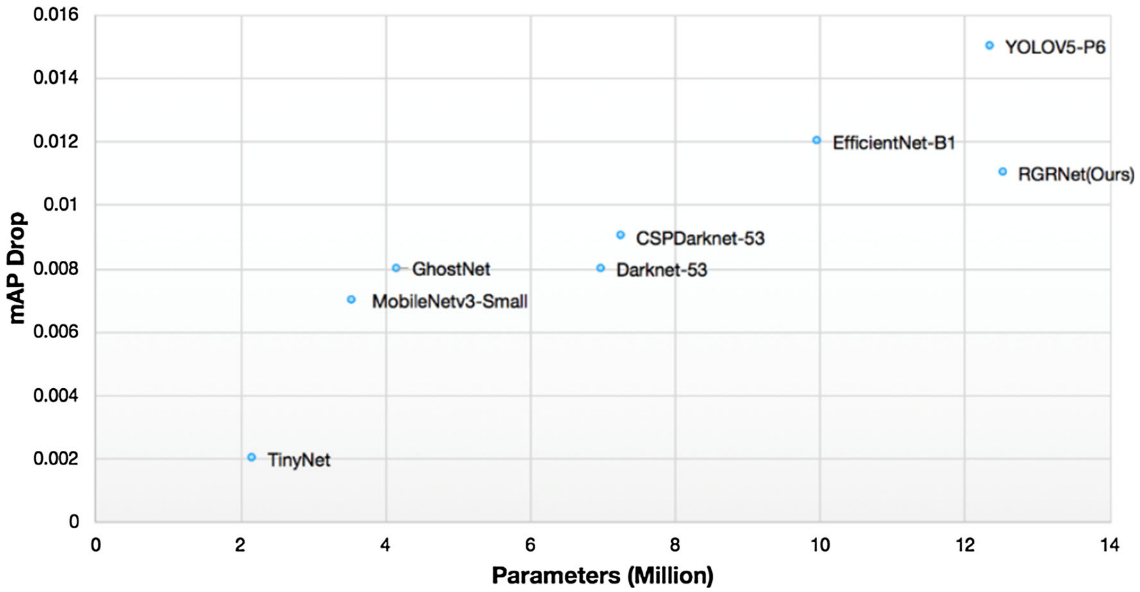


Fig. 6 Parameters vs mAP dropped on motion-blurred dataset

their weight sizes ranged from 14MB to 20MB. Medium network structures usually have a smaller weights size because some have adopted ghost convolution which reduces the number of parameters and kernel sizes in the feature extraction blocks. However, this type of network usually has lower detection accuracy too. Although we assumed smaller models would have a faster detection speed, the image process speed is surprisingly similar for CSPDarknet-53 on GTX 1660 SUPER GPU, in comparison with other lighter models. Finally, light network structures, including TinyNet and MobileNet, showed efficient inference computation with reasonable detection speed; however, their performance in terms of mAP is not promising.

5.3.2 Performance analysis on the noisy dataset comprising clear and blurred images (4:1) of hand gestures

In this experiment, we included the blurred images in the motion-blurred dataset to create a dataset with 80% clear images and 20% blurred images. As shown in Table 4, the performances of all network structures dropped, we have also highlighted the highest detection accuracy, lowest accuracy drop and fastest inference speed in bold text. The result shows our model still achieved better mAP than the other baselines.

Overall, we found that complex network structures had a higher drop in mAP; see Fig. 6. The cause of such drop is mainly due to the higher number of layers. We can see that, with more layers and parameters, these models usually have more learning capacity that encourages the negative effect of noisy data to be amplified during the learning. Although the number parameters of our model, RGRNet, are the greatest among all the models tested, the amount of accuracy dropped is still only as much as a medium-sized model.

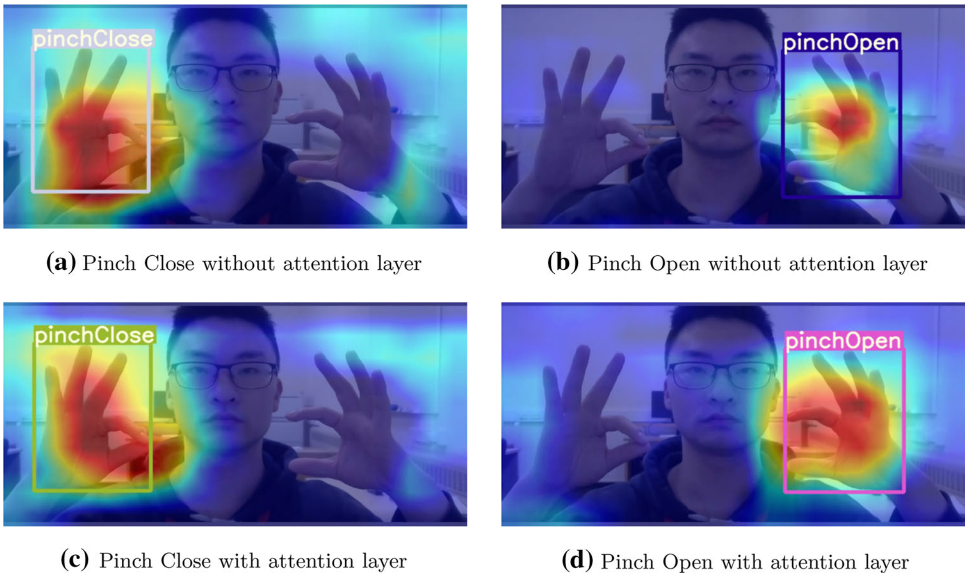
Why attention layers would work?

In the learning process of the neural network, the network generates different features to cover different semantic information. Xie et al. found that more information is beneficial to the training of the neural network [55]. By introducing an attention mechanism, our network will learn how to better capture finger-specific attention information, thus helping the network to effectively distinguish different gestures. On the other hand, transformer layers enable deep neural networks to obtain global information. The transformer layer and attention layer used in our network thus make our network able to learn both the local feature and the global feature, thereby providing greater effectiveness on similar hand gesture detection tasks. To analyse the effectiveness of our attention layers, we have printed out the attention map in Fig. 7 to highlight the important regions in the image for the detection of two similar gestures. We can see that after adding the attention layer, our model focuses on the recognition of the gesture as a whole in the detection of similar gestures, and takes into account the fingertip part in the detection of both pinch open and pinch close gestures. More importantly, we have added an extra step to analyse how attention layers would impact detection accuracy in Table 5, in bold text, we have highlighted that SELayer achieved the highest precision, recall, mAP@0.5 and mAP@0.5:0.95. Our result shows attention layers can help improve the performance where the SE layer performs better than transformer on motion-blurred dataset, which means learning local features enables our network to perform better on blurred images.

5.3.3 Similar gesture classification

In our experiment, we have analysed the performance of different networks to detect the two similar gestures on the clear dataset, ‘pinch open’ and ‘pinch close’; see Fig. 8. In Table 6, we show how different network structures performed on classifying 5 different gestures in the UTAS7k dataset, where ‘all classes’ evaluates the average detection accuracy for all gestures. In general, all models perform well on classifying ‘open’ gestures and have relatively lower detection accuracies on ‘pinch open’ gestures.

We also found that in many cases, most of the neural networks had mistaken ‘pinch close’ for ‘pinch open’. As the result, there is a higher detection accuracy for ‘pinch close’ than for ‘pinch open’, but generally this accuracy



(a) Pinch Close without attention layer

(b) Pinch Open without attention layer

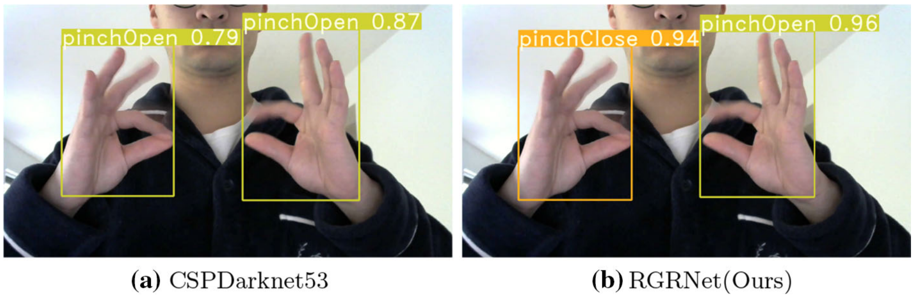
(c) Pinch Close with attention layer

(d) Pinch Open with attention layer

Fig. 7 Attention map for the models to predict similar gestures (pinch open and pinch close gestures)

Table 5 Comparison of how different attention mechanisms performed on the noisy dataset (80% clear and 20% blurred images) (image tested on GTX 1660 SUPER) classification accuracy of similar gestures

Network structure	Precision	Recall	mAP 0.5	mAP 0.5:0.95
YOLOV5-P6	0.915	0.902	0.887	0.761
YOLOV5-P6+Transformer	0.921	0.9	0.891	0.759
YOLOV5-P6 +SE layer	**0.927**	**0.914**	**0.903**	**0.765**



(a) CSPDarknet53

(b) RGRNet(Ours)

Fig. 8 Visualisation result from CSPDarknet-53 and our network, RGRNet. Note that CSPDarknet-53 incorrectly detects the fingers as 'pinch open' when the gesture is 'pinch close', see supplementary video 1

Table 6 Comparison of how accurately different networks detect UTAS7k gestures (image tested on GTX 1660 SUPER)

Network structure	All classes	Open	Close	Flip	Pinch open	Pinch close
GhostNet	0.735	0.851	0.770	0.769	0.511	0.775
CSPDarknet-53	0.753	0.876	0.788	0.796	0.541	0.761
MobileNetv3-small	0.701	0.774	0.714	0.714	0.57	0.73
TinyNet	0.703	0.800	0.825	0.711	0.542	0.739
Darknet-53	0.755	0.858	0.766	0.774	0.612	0.766
EfficientNet-B1	0.757	0.857	0.773	0.785	0.599	0.772
YOLOV5-P6	0.776	**0.863**	**0.796**	**0.769**	0.627	**0.827**
RGRNet (ours)	**0.782**	0.861	0.783	0.763	**0.685**	0.820

would not be reliable enough for most medical and neuroscience research applications as a balance error rate of these two gestures is desirable.

We can also see that adding the transformer block not only increases the overall gesture detection performance but also significantly increases the model's detection efficiency for the 'pinch open' gesture, which reduces the probability of the model misclassifying similar gestures. Moreover, results show that adding attention layers increase the total mAP for similar gestures and decreased the mAP difference between similar gestures, which also indicates that the model is more balanced in detecting similar gestures and minimise the possibility of misclassification.

6 Conclusions and future work

In this paper, we have implemented a novel network structure, RGRNet, for accurately classifying similar hand gestures and for motion-blurred image detection. Although previous methods had achieved real-time hand gesture detection, there had not been any focus on real-world fast-moving hand gesture detection in the home and clinical settings with cluttered backgrounds and ambient lighting, nor on hand detection when motion blur is present or when similar gestures are present in those blurred images. We have also developed UTAS7k, a new dataset of 7071 images (videos) with the widest variety of individual hands from 1,900 older adults and including 4: 1 clear:motion-blurred images.

We compared the detection performance of different network structures on classifying similar gesture and motion-blurred gestures, where we found multiple scale detection is effective. More importantly, our method RGRNet achieved optimising results on both similar gesture and motion-blurred gesture classification. Our assessment of a range of networks, including our new network, on these images makes a significant research contribution with a range of real-life applications. Our new dataset UTAS7k provides an important resource for the study of motion blur on hand gesture detection.

In this paper, we have shown attention mechanism is effective in classifying motion-blurred hand gestures and similar gestures. We have only used one attention module, the squeeze-and-excitation block, in our experiments, and it remains to be seen whether other attention modules will give better performance. We have used several strategies to improve the accuracy of the model and have succeeded in improving the classification accuracy for similar gestures. Essentially, we are sacrificing a portion of the speed of detection to improve performance accuracy, but still maintain real-time efficiency. Moreover, the implementation of the transformer block requires additional computing resources in the training process.

The proposed network can be embedded into a user–computer interface for clinical and neuroscience applications. It can be used for detecting different hand gestures in the hand movement tests performed by older adults, which increase the robustness of the data collection process. For future work, we will improve our model with a de-blurring attention mechanism and analyse how high resolution images would impact the inference speed of hand gesture detection. We will also investigate how complex background such as human skin-like background would impact the performance of hand gesture detection.

Supplementary information We have also included a video file named ‘Supplementary Video 1’ as the accompanying supplementary file, this file illustrates how our model detect all five gestures in UTAS7k in real time.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00521-022-08090-8.

Acknowledgements We would like to thank all the participants of The ISLAND Project who provided the video data for this study. We would also like to thank Professor James Vickers and all the staff at the University of Tasmania who work on The ISLAND Project for their support. We acknowledge the funding contributions for this project from the Medical Research Future Fund, St Lukes Health, Tasmanian Masonic Medical Research Foundation and the J.O. and J.R. Wicking Trust (Equity Trustees). We acknowledge funding from the National Health and Medical Research Council for the TAS Test Project.

Data availability The datasets generated during and/or analysed during the current study are available from the authors on reasonable request.

References

1. Alex K, Sutskever I, Hinton GE Imagenet classification with deep convolutional networks. In: NIPS’12 Proceedings of the 25th international conference on neural information processing systems, Vol. 1; pp. 1097–1105
2. LeCun Y, Boser B, Denker JS, Henderson D, Howard RE, Hubbard W, Jackel LD (1989) Backpropagation applied to handwritten zip code recognition. Neural Comput 1(4):541–551
3. Al-Hammadi M, Muhammad G, Abdul W, Alsulaiman M, Bencherif MA, Mekhtiche MA (2020) Hand gesture recognition for sign language using 3dcnn. IEEE Access 8:79491–79509
4. Zadikoff C, Lang AE (2005) Apraxia in movement disorders. Brain 128(7):1480–1497
5. Alty J, Bai Q, Li R, Lawler K, St George RJ, Hill E, Bindoff A, Garg S, Wang X, Huang G et al (2022) The TAS Test project: a prospective longitudinal validation of new online motor-cognitive tests to detect preclinical alzheimer’s disease and estimate 5-year risks of cognitive decline and dementia. BMC Neurol 22(1):1–13
6. Alty J, Bai Q, George RJS, Bindoff A, Li R, Lawler K, Hill E, Garg S, Bartlett L, King AE, Vickers JC (2021) Tastest: moving towards a digital screening test for pre-clinical Alzheimer’s disease. Alzheimer’s Dementia 17(S5):058732. https://doi.org/10.1002/alz.058732 (https://alz-journals.onlinelibrary.wiley.com/doi/pdf/10.1002/alz.058732)
7. Goetz CG, Fahn S, Martinez-Martin P, Poewe W, Sampaio C, Stebbins GT, Stern MB, Tilley BC, Dodel R, Dubois B et al (2007) Movement disorder society-sponsored revision of the unified Parkinson’s disease rating scale (mds-updrs): process, format, and clinimetric testing plan. Movement Disorders 22(1):41–47
8. Bochkovskiy A, Wang C-Y, Liao H-YM (2020) Yolov4: optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934
9. Lee M, Bae J (2020) Deep learning based real-time recognition of dynamic finger gestures using a data glove. IEEE Access 8:219923–219933. https://doi.org/10.1109/ACCESS.2020.3039401
10. Jung P-G, Lim G, Kim S, Kong K (2015) A wearable gesture recognition device for detecting muscular activities based on air-pressure sensors. IEEE Trans Ind Inf 11(2):485–494
11. Premaratne P (2014) Historical development of hand gesture recognition. Springer, Cham, pp 5–29
12. Ahmed M, Zaidan B, Zaidan A, Alamoodi A, Albahri O, Al-Qaysi Z, Albahri A, Salih MM (2021) Real-time sign language framework based on wearable device: analysis of msl, dataglove, and gesture recognition. Soft Comput, 1–22
13. Zhu Y, Yang Z, Yuan B (2013) Vision based hand gesture recognition. In: 2013 international conference on service sciences (ICSS), pp. 260–265. IEEE
14. Lee H-K, Kim J-H (1999) An hmm-based threshold model approach for gesture recognition. IEEE Trans Pattern Anal Mach Intell 21(10):961–973
15. Marcel S, Bernier O, Viallet J-E, Collobert D (2000) Hand gesture recognition using input-output hidden Markov models. In: proceedings fourth IEEE international conference on automatic face and gesture recognition (Cat. No. PR00580), pp. 456–461. IEEE
16. Ng CW, Ranganath S (2002) Real-time gesture recognition system and application. Image Vis Comput 20(13–14):993–1007
17. Chen Q, Georganas ND, Petriu EM (2008) Hand gesture recognition using haar-like features and a stochastic context-free grammar. IEEE Trans Instrum Meas 57(8):1562–1571
18. Mohanty A, Rambhatla SS, Sahay RR (2017) Deep gesture: static hand gesture recognition using CNN. In: proceedings of international conference on computer vision and image processing, pp. 449–461. Springer
19. Bose SR, Kumar VS (2020) Efficient inception v2 based deep convolutional neural network for real-time hand action recognition. IET Image Process 14(4):688–696
20. Yi C, Zhou L, Wang Z, Sun Z, Tan C (2018) Long-range hand gesture recognition with joint ssd network. In: 2018 IEEE international conference on robotics and biomimetics (ROBIO), pp. 1959–1963. IEEE
21. Mujahid A, Awan MJ, Yasin A, Mohammed MA, Damaševičius R, Maskeliūnas R, Abdulkareem KH (2021) Real-time hand gesture recognition based on deep learning yolov3 model. Appl Sci 11(9):4164
22. Benitez-Garcia G, Prudente-Tixteco L, Castro-Madrid LC, Toscano-Medina R, Olivares-Mercado J, Sanchez-Perez G, Villalba LJG (2021) Improving real-time hand gesture recognition with semantic segmentation. Sensors 21(2):356
23. Benitez-Garcia G, Olivares-Mercado J, Sanchez-Perez G, Yanai K (2021) IPN hand: a video dataset and benchmark for real-time continuous hand gesture recognition. In: 2020 25th international conference on pattern recognition (ICPR), pp. 4340–4347. IEEE
24. Gupta P, Kautz K, et al (2016) Online detection and classification of dynamic hand gestures with recurrent 3d convolutional neural networks. In: CVPR, vol 1, p. 3
25. Köpüklü O, Gunduz A, Kose N, Rigoll G (2019) Real-time hand gesture detection and classification using convolutional neural networks. In: 2019 14th IEEE international conference on automatic face & gesture recognition (FG 2019), pp. 1–8. IEEE

26. Do N-T, Kim S-H, Yang H-J, Lee G-S (2020) Robust hand shape features for dynamic hand gesture recognition using multi-level feature lstm. Appl Sci 10(18):6293
27. Redmon J, Divvala S, Girshick R, Farhadi A (2016) You only look once: unified, real-time object detection. In: proceedings of the IEEE conference on computer vision and pattern recognition, pp. 779–788
28. Ni Z, Chen J, Sang N, Gao C, Liu L (2018) Light yolo for high-speed gesture recognition. In: 2018 25th IEEE international conference on image processing (ICIP), pp. 3099–3103. IEEE
29. Redmon J, Farhadi A (2017) Yolo9000: better, faster, stronger. In: proceedings of the IEEE conference on computer vision and pattern recognition, pp. 7263–7271
30. Redmon J, Farhadi A (2018) Yolov3: an incremental improvement. arXiv preprint arXiv:1804.02767
31. Jocher G, et al. (2021) ultralytics/yolov5: V5.0 - YOLOv5-P6 1280 Models, AWS, Supervise.ly and YouTube integrations. https://doi.org/10.5281/zenodo.4679653
32. Xianbao C, Guihua Q, Yu J, Zhaomin Z (2021) An improved small object detection method based on yolo v3. Pattern Anal Appl 1–9
33. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778
34. Ross T-Y, Dollár G (2017) Focal loss for dense object detection. In: proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2980–2988
35. Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu C-Y, Berg AC (2016) SSD: single shot multibox detector. In: European conference on computer vision, pp. 21–37. Springer
36. Tan M, Le Q (2019) Efficientnet: rethinking model scaling for convolutional neural networks. In: international conference on machine learning, pp. 6105–6114. PMLR
37. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, et al (2020) An image is worth 16x16 words: transformers for image recognition at scale. arXiv preprint arXiv:2010.11929
38. Wang C-Y, Liao H-YM, Wu Y-H, Chen P-Y, Hsieh J-W, Yeh I-H (2020) CSPNet: a new backbone that can enhance learning capability of CNN. In: proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops, pp. 390–391
39. Lin T-Y, Dollár P, Girshick R, He K, Hariharan B, Belongie S (2017) Feature pyramid networks for object detection. In: proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2117–2125
40. Wang K, Liew JH, Zou Y, Zhou D, Feng J (2019) Panet: few-shot image semantic segmentation with prototype alignment. In: proceedings of the IEEE/CVF international conference on computer vision, pp. 9197–9206
41. Ridnik T, Lawen H, Noy A, Ben Baruch E, Sharir G, Friedman I (2021) TRESNet: high performance GPU-dedicated architecture. In: proceedings of the IEEE/CVF winter conference on applications of computer vision, pp. 1400–1409
42. Elfwing S, Uchibe E, Doya K (2018) Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. Neural Netw 107:3–11
43. Hu J, Shen L, Sun G (2018) Squeeze-and-excitation networks. In: proceedings of the IEEE conference on computer vision and pattern recognition, pp. 7132–7141
44. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. Adv Neural Inf Process Syst 30:5998–6008
45. Bartlett L, Doherty K, Farrow M, Kim S, Hill E, King A, Alty J, Eccleston C, Kitsos A, Bindoff A et al (2022) Island study linking aging and neurodegenerative disease (island) targeting dementia risk reduction: protocol for a prospective web-based cohort study. JMIR Res Protoc 11(3):34688
46. Afifi M (2019) 11k hands: gender recognition and biometric identification using a large dataset of hand images. Multimed Tools Appl. https://doi.org/10.1007/s11042-019-7424-8
47. Sun Z, Tan T, Wang Y, Li S (2005) Ordinal palmprint representation for personal identification. In: proceedings of the IEEE conference on computer vision and pattern recognition
48. Abdesselam A, Al-Busaidi A (2012) Person identification prototype using hand geometry. https://doi.org/10.13140/2.1.2181.9844
49. Kumar A (2008) Incorporating cohort information for reliable palmprint authentication. In: 2008 Sixth Indian conference on computer vision, graphics & image processing, pp. 583–590. IEEE
50. Ferrer MA, Morales A, Travieso CM, Alonso JB (2007) Low cost multimodal biometric identification system based on hand geometry, palm and finger print texture. In: 2007 41st annual IEEE international Carnahan conference on security technology, pp. 52–58. IEEE
51. Pech-Pacheco JL, Cristóbal G, Chamorro-Martinez J, Fernández-Valdivia J (2000) Diatom autofocusing in brightfield microscopy: a comparative study. In: proceedings 15th international conference on pattern recognition. ICPR-2000, vol. 3, pp. 314–317. IEEE
52. Han K, Wang Y, Tian Q, Guo J, Xu C, Xu C (2020) GhostNet: more features from cheap operations
53. Howard A, Sandler M, Chu G, Chen L-C, Chen B, Tan M, Wang W, Zhu Y, Pang R, Vasudevan V, Le QV, Adam H (2019) Searching for MobileNetV3
54. Lin T-Y, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL (2014) Microsoft coco: common objects in context. In: European conference on computer vision, pp. 740–755. Springer
55. Xie T, Deng J, Cheng X, Liu M, Wang X, Liu M (2022) Feature mining: a novel training strategy for convolutional neural network. Appl Sci 12(7):3318