Displays 77 (2023) 102388

Contents lists available at ScienceDirect

Displays

journal homepage: www.elsevier.com/locate/displa

DeReFNet: Dual-stream Dense Residual Fusion Network for static hand gesture recognition☆

Jaya Prakash Sahoo [a,*], Suraj Prakash Sahoo [b], Samit Ari [a], Sarat Kumar Patra [c]

[a] *Department of Electronics and Communication Engineering, National Institute of Technology Rourkela, Odisha, 769008, India*
[b] *School of Electronics Engineering, Vellore Institute of Technology Vellore, Tamil Nadu, 632014, India*
[c] *Indian Institute of Information Technology Vadodara, Gujarat, 382028, India*

ARTICLE INFO

Keywords:
Convolutional neural network
Dense connectivity
Feature concatenation module
Global feature aggregation block
Hand gesture recognition
Residual network

ABSTRACT

Vision-based hand gesture recognition (HGR) system provides the most effective and natural way of interaction between humans and machines. However, the recognition performance of such an HGR system is challenging due to the variations in illumination, complex backgrounds, the shape of the user's hand, and inter-class similarity. This work proposes a compact dual-stream dense residual fusion network (DeReFNet) to address the above challenges. The proposed convolutional neural network architecture mainly utilizes the strength of global features from each residual block of the residual stream and spatial information from the other stream using dense connectivity. Both the streams are fused to gather enriched information using the feature concatenation module. The efficacy of the DeReFNet is validated using a subject-independent cross-validation technique on four publicly available benchmark datasets. Furthermore, the qualitative and quantitative analysis of the benchmarked datasets illustrates that the DeReFNet outperforms state-of-the-art methods in terms of accuracy and computational time.

1. Introduction

In society, hand gestures are generally used as a non-verbal communication medium for deaf and dumb people to convey their information [1]. Due to its friendliness and flexibility, it is also applied to develop a human–computer interface and human–robot interaction systems. Several potential applications of hand gesture-based systems reported in the literature are real-time automotive interface [2], sign language interpretation [3], and gaming [4] etc. Therefore, the interest and need for a gesture interpretation system motivate the researchers to develop an accurate hand gesture recognition system.

Substantial works have been performed over the last several decades on glove-based and vision-based hand gesture recognition (HGR) techniques. This literature [3,5] justify that the vision-based technique is superior and user-friendly over the glove-based technique due to the following reasons (i) it does not require external hardware attached to the user's hand, (ii) cost reduction, (iii) gestures are recognized using the computer vision techniques. In vision-based techniques, gestures are recognized using the following steps: pre-processing, feature extraction, and classification [3,5,6]. However, the gesture recognition performance is affected by the challenges such as illumination variation, hand segmentation, hand shape, and inter-class similarity gesture poses. Several approaches have been proposed in the literature to solve the above difficulties. For illumination normalization, homomorphic filtering and grey world methods are suggested [7]. For hand region segmentation simple threshold technique is applied on uniform backgrounds [8,9] and skin colour-based filtering techniques [3,10] are used to overcome the skin colour noise. However, hand segmentation is still difficult when it is surrounded by any skin colour objects [3] and complex backgrounds. In most of the literature, the reported performance is good if the number of gesture classes is less in the dataset. However, for a large number of gesture class such as American sign language (ASL), the gesture recognition performance is significantly reduced using the reported techniques [4]. The reason for this is ASL gesture images have very significant inter-class similarities. Most standard feature extraction algorithms frequently miss essential information to differentiate similar gesture postures in the dataset.

After the segmentation of the hand region, some researchers used a hand-crafted feature extraction technique followed by gesture classification for HGR. Similarly, deep learning network such as convolutional neural network (CNN) is also developed to perform both tasks in a single network. Several hand crafted feature extraction techniques proposed in the literature are discrete wavelet transform (DWT) and Fisher

☆ This paper was recommended for publication by Menghan Hu.
* Corresponding author.
E-mail addresses: sahoo.jprakash@gmail.com (J.P. Sahoo), samit@nitrkl.ac.in (S. Ari).

https://doi.org/10.1016/j.displa.2023.102388
Received 2 August 2022; Received in revised form 6 December 2022; Accepted 25 January 2023
Available online 6 February 2023
0141-9382/© 2023 Published by Elsevier B.V.

ratio (F-ratio) [3], contour features [11–13], etc. These attributes work effectively in a specified context and have poor generalization properties in various situations [14]. On the other hand, deep learning techniques such as CNN architectures are adapted to overcome the above limitation by various researchers [15,16]. Among the CNN architectures, two widely used networks are residual network (ResNet) [17], and densely connected network (DenseNet) [18] which are mostly used to overcome the gradient vanishing problem and parameter efficiency, respectively. In ResNet, an identity skip connection is provided in the residual block to alleviate the vanishing of the gradient problem. For feature propagation and gradient flow, dense connectivity is provided within the dense block of DenseNet. Even though both architectures perform well and are very popular, they have specific limitations. For ResNet, the identity shortcut bypasses the residual blocks to conserve features. This skip connection limits the network's full representational potential, leading to the collapsing domain issue and reducing its learning ability [19]. DenseNet utilizes dense concatenation to all subsequent layers to avoid direct summing, preserving preceding layer information. DenseNet is proven to have a better feature utilization efficiency than ResNet with less number of parameters [18]. However, DenseNet requires large GPU memory due to feature concatenation, and it also takes more training time. Therefore, there is a need for a compact CNN architecture with fewer trainable parameters and less average inference time in the field of HGR.

The work of this paper proposes a compact dual-stream fusion network to recognize vision-based hand gestures effectively. The proposed architecture consists of GFA residual stream, SF dense stream, and a feature concatenation module (FCM). Through global average pooling technique, the GFA residual stream aims to extract low, mid, and high-level features from hand gesture images. Then, all these information are combined to represent the gesture attributes efficiently. The SF dense stream combines the spatial information of gesture images through feature reuse, which strengthens the network by extracting the refined local-to-global texture features of hand gesture images. The combination of feature elements from both the streams using FCM can predict the gesture class accurately even in the presence of complex backgrounds and human skin coloured objects The contributions of this paper are as follows:

- An end-to-end dual-stream dense residual fusion network (DeReFNet) is proposed to recognize hand gestures accurately. One stream in the DeReFNet is a global feature aggregation block-based (GFA) residual stream, while the other is a collection of spatial features (SF) through dense connectivity, i.e., SF dense stream. The fusion of both the streams can distinguish the inter-class similarity gestures in the gesture datasets.
- The developed GFA residual stream is intended to capture the deep knowledge of hand gestures by the fusion of global information from each residual block in the residual stream.
- The proposed DeReFNet is designed with less trainable parameters, making the network computationally efficient with less average inference time than other existing CNNs.
- Extensive experiments are performed on four challenging vision-based static hand gesture datasets, proving that the proposed DeReFNet outperforms the existing techniques in terms of mean accuracy and computational time. Detailed qualitative and quantitative analysis is also presented to provide insight into the proposed DeReFNet.

The rest of the paper is organized as follows. A review of related literature on HGR is provided in Section 2. The proposed methodology is discussed in Section 3. Section 4 describes the extensive experimental results and discussions to validate the proposed technique. The conclusions of the paper are given in Section 5.

Table 1
Summary of literature survey on static hand gesture recognition techniques.

Literature	Year	Proposed work	Solutions
Fang et al. [11]	2020	Hand geometric features	Optimized shape features from hand contour for HGR
Lazarou et al. [12]	2021	Angular-radial bins (ARB) based shape descriptors	Hand shape descriptors form binary image are obtained to represent a gesture image
Wang [13]	2021	SDD featuresform hand contour	Sparse representation of SDD and matching algorithm for HGR
Tan et al. [14]	2021	EDenseNet	Modified transition layer in typical DenseNet architecture
Zhou et al. [15]	2022	A two-stage HGR system	CNN architecture for learning of features from segmented and RGB image using two different channel
Sahoo et al. [20]	2022	RBI2RCNN	Residual block intensity (RBI) features from 2RCNN with SVM classifier
Bhaumik et al. [16]	2021	ExtriDeNet	Fusion of multi-scale features in a CNN for HGR
Chevtchenko et al. [21]	2018	FFCNN	Fusion of Gabor feature with CNN at FC layer
Sahoo et al. [22]	2022	Fusion of deep features	Combination of FC layer feature from AlexNet and VGG 16
Proposed	–	DeReFNet	A dual-stream dense residual fusion network to reduce the inter class gesture poses in the dataset

2. Related works

A complete literature survey is provided in this section to illustrate the most current known contribution in the vision-based static HGR domain and its limitations (see Table 1).

The hand-crafted-based approach was superior to all other methods for recognizing hand gesture images before the deep learning technique appeared. This approach generally includes different steps of pre-processing operations to extract the hand region from the image frame. Then, the features are obtained from the pre-processed image using the hand-crafted features and recognized using different classifiers. Fang et al. [11] provided new hand shape descriptors for the recognition of hand postures. A local descriptor is constructed by extracting several geometric feature vectors from the segmented hand's contour, such as distance, angle, and curvature. Then, the features are optimized using a Fisher vector to represent the hand shape, and the gesture classes are recognized using a multi-class SVM classifier. Novel shape descriptors known as angular-radial bins are introduced by the Lazarou et al. [12] for recognition of hand gestures using the shape matching algorithm. Wang [13] proposed slope difference distribution (SDD) features, representing the distance between the hand centroid and each point in the hand contour defined in sparse. Then gesture classes are recognized using the model matching algorithm. However, hand-crafted characteristics are influenced by hand form distortion caused by incorrect hand segmentation, light variability, dynamic backgrounds, and so on [11]. Moreover, these features are advantageous for addressing specific problems in a dataset but exhibit poor generalization behaviour for various situations in the datasets.

Nowadays, deep learning [14,16] approaches offer better performance in visual analysis and classification because they can automatically learn useful features [22,23] from the input images. enhanced

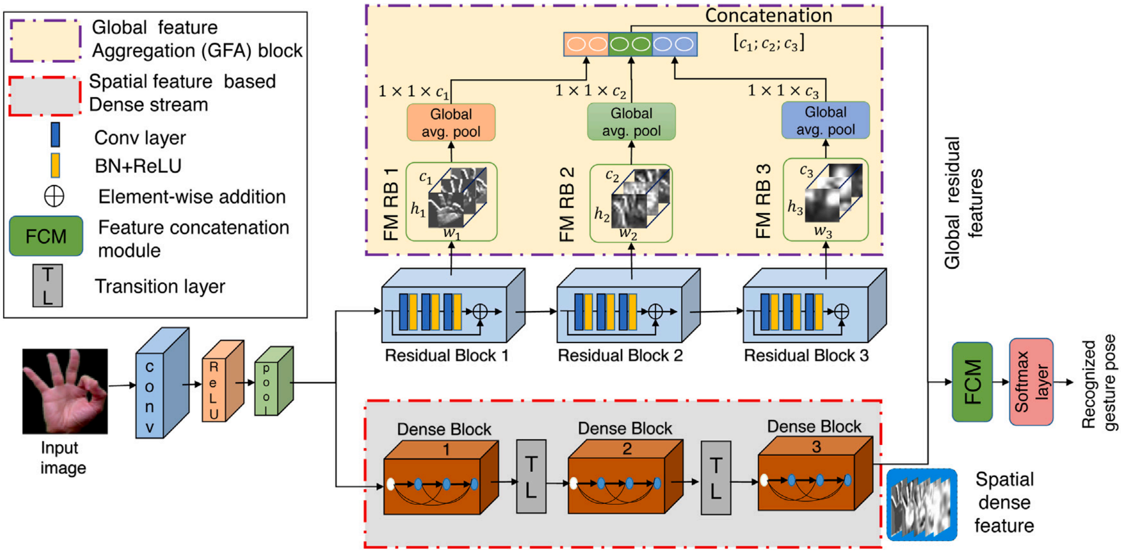

Fig. 1. Overview of the proposed hand gesture recognition system. Here, FM: feature maps, RB: residual blocks.

densely connected convolutional neural network (EDenseNet) is proposed by Tan et al. [14] for recognition of hand gestures. The authors modified the transition layers in a typical DenseNet architecture in the proposed architecture to strengthen the feature propagation. A two-stage HGR system is presented by Zhou et al. [15] for hand gesture recognition in complex backgrounds. The authors proposed a hand segmentation network based on a novel encoder–decoder architecture with dilated residual network and atrous spatial pyramid pooling module. A double-channel CNNs architecture is developed, which learns features from the RGB input images and the segmented hand images separately. A residual block intensity (RBI) features using a two-stage residual CNN (2RCNN) is reported by Sahoo et al. [20] for accurate recognition of hand gestures in the presence of complex backgrounds and human skin colour noise. The reported RBI features explore the global and local information obtained from each receptive field of 2RCNN. A light weighted intensive feature extrication deep network (ExtriDeNet) is proposed by Bhaumik et al. [16] for recognition of hand gesture. In the reported work, the multi-scale features are obtained from the hand postures to represent the proposed features. Although deep learning has excellent generalization capabilities, it demands a completely developed CNN architecture and large storage capacity.

Several researchers have also proposed the fusion of hand-crafted features with CNN architecture and the fusion of deep features from the fully connected (FC) layers of the pre-trained deep CNNs. A feature fusion-based CNN (FFCNN) architecture is proposed by Chevtchenko et al. [21] for the recognition of hand postures. In this reported technique, the authors fused the hand-crafted Gabor features with the CNN architecture at fully connected layers to recognize hand postures. Fusion of features from FC layer of the pre-trained CNNs are proposed by Sahoo et al. [22] for recognizing hand gestures. The authors found that concatenating features, particularly the ‘FC6’ layer of pre-trained AlexNet and VGG 16, outperform individual FC layer features regarding recognition accuracy. In another work, the authors [23] proposed the optimized features of ‘FC6’ layers of pre-trained AlexNet for HGR using the SVM classifier. The authors found that the optimized deep features using the principal component analysis technique outperformed the individual FC layer features. Moreover, the existing HGR technique has large parameters and thus requires more computational time to recognize a hand gesture.

To overcome the above issue, the proposed work intends to develop a compact dual-stream CNN architecture that extracts the global and spatial features using two different streams to recognize hand gestures effectively.

3. Proposed methodology

The block diagram representation of the proposed HGR system using DeReFNet is illustrated in Fig. 1. In this network, one stream is a global feature aggregation block-based residual stream for learning global features, and the other stream is a spatial feature-based dense stream. The fusion of both streams using the feature concatenation module (FCM) develops the end-to-end network known as DeReFNet.

3.1. Global feature aggregation (GFA) block-based residual stream

In CNN, several layers are stacked together to develop a deeper neural network. The deeper layers can learn more discriminating features from the input images. However, training using the deeper CNN is challenging due to the saturation of accuracy and degradation of gradients in the developed CNN. To solve the above gradient vanishing issue, He et al. [17] proposed a residual learning technique with residual blocks in the CNN architecture. Identity skip connections are used in this learning approach, which permits gradient flow from top to bottom. The structure of the residual unit used in this work is shown in Fig. 2. The residual units in this network have the typical connection of Conv, BN, and ReLU layer followed by an addition from the residual block’s input. Each residual unit can be defined as:

$$y_r = F(x_r, W_r) + x_r \tag{1}$$

where x_r and y_r indicate the input and output of the r^{th} unit and the residual function F is defined by the weight W.

The global pooling descriptor takes the output from each residual blocks i.e., fm_c which has size $h \times w$ with c feature maps. The global average pooling descriptor shrinks the size from $h \times w \times c$ to $1 \times 1 \times c$ after the processing as:

$$g_p = \frac{1}{h \times w} \sum_{i=1}^{h} \sum_{j=1}^{w} fm_c(i, j) \tag{2}$$

where $fm_c(i, j)$ is the value at location (i, j) in the FM. The g_p features obtain from residual block 1, 2, and 3 generate the feature of dimension $1 \times 1 \times c_1$, $1 \times 1 \times c_2$, and $1 \times 1 \times c_3$ respectively. Then each g_p are concatenated to represent the global feature aggregation block for the residual stream (g_{rf}) represented as:

$$g_{rf} = [c_1; c_2; c_3] \tag{3}$$

The final output g_{rf} is given as one of the inputs to the FCM.

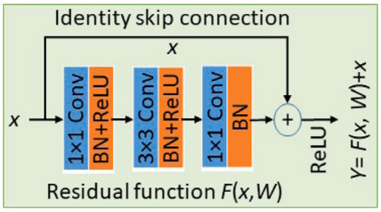

Fig. 2. Structure of residual unit used in the residual stream. Here, BN: batch normalization, ReLU: rectified linear unit, Conv: convolutional layer.

Table 2
Layer-wise details of the global feature aggregation block-based residual stream of the DeReFNet.

Layer name	Output size	Layer
Input	$64 \times 64 \times 3$	–
Convolution	$64 \times 64 \times 64$	3×3, f, stride 1
ReLU	$64 \times 64 \times 64$	–
Pooling	$32 \times 32 \times 64$	2×2, max pool, stride 2
Residual Block 1 (res1)	$32 \times 32 \times 128$	$\begin{bmatrix} 1 \times 1, 64 \\ 3 \times 3, 64 \\ 1 \times 1, 128 \end{bmatrix} \times 2$
Residual Block 2 (res2)	$16 \times 16 \times 256$	$\begin{bmatrix} 1 \times 1, 128 \\ 3 \times 3, 128 \\ 1 \times 1, 256 \end{bmatrix} \times 2$, stride 2
Residual Block 3 (res3)	$8 \times 8 \times 512$	$\begin{bmatrix} 1 \times 1, 256 \\ 3 \times 3, 256 \\ 1 \times 1, 512 \end{bmatrix} \times 2$, stride 2
global pool res3	$1 \times 1 \times 512$	Global avg. pool
global pool res2	$1 \times 1 \times 256$	Global avg. pool
Depth concat_1 [res3, res2]	$1 \times 1 \times 768$	Depth concatenation
global pool res1	$1 \times 1 \times 128$	Global avg. pooling
Depth concat_2 [res1, Depth concat_1]	$1 \times 1 \times 896$	Depth concatenation
FCM input-1	$1 \times 1 \times 896$	Depth concatenation

Table 3
Layer-wise details of the spatial feature-based dense stream of the proposed DeReFNet with growth rate of k.

Layer	Output size	Layer
Input	$64 \times 64 \times 3$	–
Convolution	$64 \times 64 \times 64$	3×3, f, stride 1
ReLU	$64 \times 64 \times 64$	–
Pooling	$32 \times 32 \times 64$	2×2, max pool, stride 2
Dense Block 1	$32 \times 32 \times (64 + 4 \times k)$	$\begin{bmatrix} \text{conv\_1} \times 1 \\ \text{conv\_3} \times 3 \end{bmatrix} \times 4$
Transition Layer 1	$32 \times 32 \times (64 + 4 \times k)$ $16 \times 16 \times (64 + 4 \times k)$	1×1 conv 2×2 Avg. pooling, stride 2
Dense Block 2	$16 \times 16 \times (64 + 8 \times k)$	$\begin{bmatrix} \text{conv\_1} \times 1 \\ \text{conv\_3} \times 3 \end{bmatrix} \times 4$
Transition Layer 2	$16 \times 16 \times (64 + 8 \times k)$ $8 \times 8 \times (64 + 8 \times k)$	1×1 conv 2×2 Avg. pooling, stride 2
Dense Block 3	$8 \times 8 \times (64 + 16 \times k)$	$\begin{bmatrix} \text{conv\_1} \times 1 \\ \text{conv\_3} \times 3 \end{bmatrix} \times 8$
global pool	$1 \times 1 \times (64 + 16 \times k)$	Global avg. pool
FCM input-2	$1 \times 1 \times (64 + 16 \times k)$	Depth concatenation

The layer-wise architecture of the GFA residual stream of the DeReFNet is presented in Table 2. In this stream, there are three residual blocks, each with two residual units. The global features from each residual block are concatenated and given as one of the inputs of the FCM.

3.2. Spatial feature-based dense stream

DenseNet uses the feature reuse concept through dense connectivity among the convolutional (conv) layers in a given dense block [18]. That means the output FM of the present conv layer is concatenated with the FM of the previous conv layer. In this way, DenseNet improves the gradient flow and feature propagation among the conv layers in a dense block. The advantages of this network are (i) it reduces the gradient vanishing problems, (ii) strengthens the feature propagation, (iii) encourages the feature reuse, (iv) reduces the number of trainable parameters.

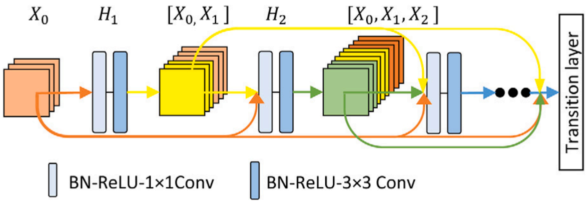

Fig. 3. Visualization of connection between the conv layers within a dense block. Each layer takes the FM from the previous layer as an input and generates a concatenation of FM using the composite function.

Dense block: In ResNet [17], a fixed number of output FM are generated from the conv layer within the residual block. Then, the FM of the first conv layer is added with the second layer FM and so on using the identity function. Thus this network improves the gradient flow and feature propagation using the summation function. But, in DenseNet, the FM is concatenated from all the preceding conv layers in the dense block. Thus, this network improves the flow of information and gradient among the conv layers. The connection of conv layers within a dense block in DenseNet is shown in Fig. 3.

The i^{th} conv layer in the dense block, receives the concatenation of FM of all the previous layers represented as $X_0, X_1, X_2, \ldots, X_{i-1}$ as input:

$$X_i = H_i([X_0, X_1, X_2, \ldots, X_{i-1}]) \tag{4}$$

where, $[X_0, X_1, X_2, \ldots, X_{i-1}]$ denotes the concatenation of FM by the previous layers $0, 1, 2, \ldots, i-1$. $H_i(.)$ is known as composite function. This function is defined using three consecutive operations such as BN, ReLU, and convolution.

Growth rate: In the network, each $H_i(.)$ produces k number of FM. Thus, i^{th} conv layer has $k_0 + k \times (i-1)$ input FM, where k_0 is the number of channels at the input of dense block. This hyper parameter k is known as growth rate. The value of k is acquired experimentally and are explained in the next section.

Transition layer: This layer is made up of 1×1 conv layer followed by an average pooling layer. This layer reduces the model complexity.

The layer-wise architecture of the spatial feature-based dense stream of the DeReFNet is presented in Table 3. In this stream, three dense blocks are denoted as dense block 1, 2, and 3 with 4, 4, and 8 dense connectivity layer structures, respectively. The spatial features from the last dense blocks are given as another input of the FCM after the pooling operations.

3.3. Feature concatenation module (FCM)

In the proposed DeReFNet, the global and spatial information from two different streams is concatenated at the Feature concatenation module (FCM). The FCM gathers the distinguishable features from the above two streams. The FM generated at different residual blocks, and dense blocks for an input gesture are shown in Fig. 4. The initial layers of the network capture the low-level features such as edges, blobs, etc., and the deeper layers represent the high-level semantic features for the input image. Finally, the FCM combines the distinct characteristics of the two streams. For example, in the GFA stream, the output feature map size in residual blocks 1, 2, and 3 are $32 \times 32 \times 128$, $16 \times 16 \times 256$, and $8 \times 8 \times 512$ respectively. Therefore the dimension of feature vectors after global average pooling at residual blocks 1, 2, and 3 are 128, 256, and 512, respectively. The feature of the GFA stream is then generated by concatenating all of the features, and its dimension is 896.

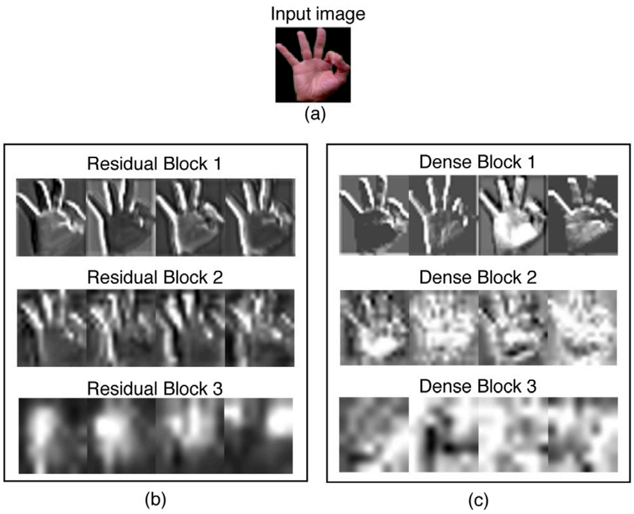

Fig. 4. Visualization of feature maps output generated at different residual and dense blocks of the DeReFNet. (a) input image, (b) FM output from three residual blocks, (c) FM outputs from three dense blocks.

Similarly, in the SF stream, the output feature dimension is 64+(16×k), where k is the growth rate and its value is obtained experimentally. For an example, k is chosen as 16 for the MU-ASL-digit dataset, therefore, the dimension of the SF stream is 320 and the final feature dimension of the FCM is 1216.

4. Experimental results and discussions

In this section, experiments are carried out to validate the datasets, various parameters study, and performance analysis of the proposed CNN.

4.1. Datasets, experimental setup, and validation methods

4.1.1. Benchmarked datasets

In this section, four benchmarked vision-based static hand gesture datasets with several challenges are chosen to evaluate the performance of the proposed method. The detailed description of the datasets are given below.

The first dataset is known as Jochen-Triesch ASL dataset [8] (denoted as JT-ASL) of 10 ASL gesture poses as shown in Fig. 5(a). The challenges in this dataset are variations in the size and shape of the gesture pose. The second dataset is known as the Massey University (MU) ASL dataset [9] and a benchmarked dataset of ASL gesture pose. The datasets consists of 10 ASL digit ‘0’-‘9’ (referred to as MU-ASL-digit) and 26 ASL alphabets ‘a’-‘z’ (denoted as MU-ASL-alphabet) gesture poses as shown in Fig. 5(b) and Fig. 5(c) respectively. The challenges in the database include variations in light from five different angles, including left, right, bottom, top, and diffuse, as well as a large range of gesture pose hand sizes. The third dataset is known as the National University of Singapore (NUS) hand posture dataset II [10] of 10 gesture poses as shown Fig. 5(d). The challenges in this dataset are variation in non-uniform backgrounds and the shape of the user’s hand. The last dataset is known as the American sign language finger spelling [24] dataset (referred to as ASL-FS-colour) is a 24 ASL alphabet (except letters ‘j’ and ‘z’ due to dynamic). The major challenge in this dataset is human skin colour noise, as shown in Fig. 5(e). The summary of all the datasets is presented in Table 4.

Table 4
Details of the datasets used to evaluate the performance of the proposed technique.

Dataset	No. of subjects	Gesture classes	Total samples
JT-ASL [8]	24	10	240
MU-ASL-digit [9]	5	10	700
MU-ASL-alphabet [9]	5	26	1815
NUS-II [10]	40	10	2000
ASL-FS-colour [24]	5	24	>60000

4.1.2. Experimental setup

The experiments are carried out in MATLAB 2019B platform using an 8 GB NVIDIA graphics card. Stochastic gradient descent with momentum (SGDM) is used as an optimization function of training. The following hyper-parameters are set for the CNN training: maximum epoch = 50, initial learning rate = 0.01, momentum = 0.9.

4.1.3. Validation methods

The performance of DeReFNet is evaluated using the leave-p-subject-out (LpO) cross-validation (CV) test. In this technique, for a dataset having S subjects, gesture samples of $S-p$ subjects are used for training, and the performance of the trained model is evaluated using gesture samples of the remaining p subjects. The process is repeated for each combination of p subjects to obtain the recognition performance in mean accuracy. The leave-one-subject-out CV (LOO CV) [16] test is conducted on all benchmarked datasets in this work.

4.2. Evaluation of the proposed technique

In this subsection, the performance of the proposed DeReFNet on the variation of growth rate and batch size is analysed. In addition, ablation studies and qualitative and quantitative analyses are also performed to validate the proposed technique.

4.2.1. Effect of performance on various parameters

Batch size variations: The hyper-parameter batch size is defined by the number of images taken in a batch as input from the training set. If the batch size is small, the CNN will converge quickly, and vice versa. The impact of batch size on model performance is shown in Fig. 6(a). It is observed from the experimental result that the system achieves the best performance on a batch size of 8 for JT-ASL and MU-ASL-digit datasets. Similarly, for MU-ASL-alphabet, NUS-II, and ASL-FS-colour-colour datasets, the performance is superior for a batch size of 16.

Variation in growth rate: The growth rate is the number of feature maps added to the global state by each convolutional layer in the dense block of the network. In this study, the performance of all the bench-marked datasets is evaluated for four different values of the growth rate, such as 8, 16, 32, and 64. The influence of performance on variation of growth rate is shown in Fig. 6(b). The figure shows that for growth rate 8, JT-ASL and MU-ASL-alphabet datasets depict superior performance. Similarly, for MU-ASL-digit, NUS-II and ASL-FS-Colour datasets, the performance is superior for the growth rate of 16.

The training and loss curves for five subject-independent LOO CV observations on the MU-ASL-alphabet dataset are shown in Fig. 6(c). The curves indicate that the model was correctly trained after proper hyper-parameter settings.

4.2.2. Ablation study

An ablation study is performed to evaluate the effectiveness of the proposed DeReFNet using the different modules. Comparisons of performance in terms of mean accuracy (%) are acquired using the ASL-FS-colour dataset. The network consists of modules such as residual stream, global feature aggregation (GFA) block, and densely connected stream. In this work, a three-stage residual stream and the dense stream

Table 5
Ablation study for the effectiveness comparison of the proposed network with different modules on ASL-FS-colour dataset.

Residual stream	✓		✓	✓	✓	✓	✓
Dense stream		✓			✓	✓	✓
GFA block (with max pooling)			✓			✓	
GFA block (with average pooling)				✓			✓
Mean accuracy (%)	86.38	86.95	87.64	88.45	88.68	90.32	**91.24**

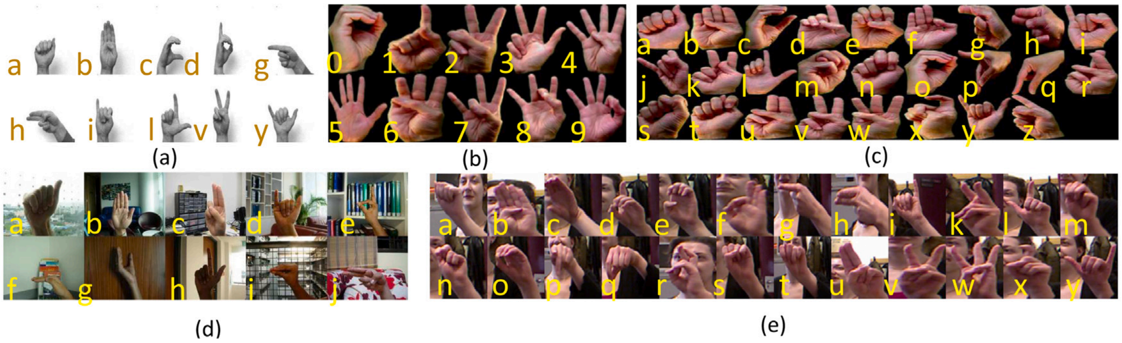

Fig. 5. Four benchmark datasets used to evaluates the performance. (a) 10 ASL gesture poses of JT-ASL, (b) 10 ASL digit ('0' - '9') gesture poses of MU-ASL-digit, (c) 26 ASL alphabet ('a' - 'z') gesture poses of MU-ASL-alphabet, and (d) 10 gesture poses of NUS-II, and (e) 24 ASL alphabet of ASL-FS-colour dataset.

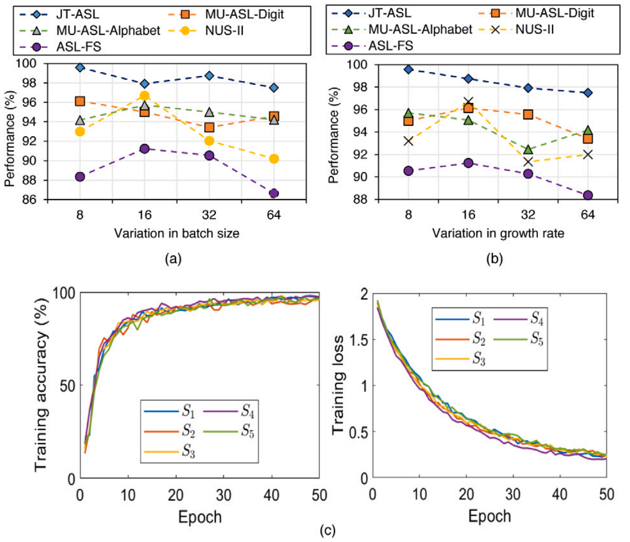

Fig. 6. The effect of DeReFNet performance on the variation of several parameters. (a) variation in batch size, (b) variation in growth rate, and (c) Training accuracy and loss curve of the proposed network on MU-ASL-alphabet dataset (where S_i (i = 1,2,...,5) represents the observation using other than ith subject data for the training of model).

are empirically chosen to develop the network architecture. Again, in GFA block, global average pooling and max pooling techniques are separately used to find the effectiveness of the proposed network. The performance in terms of mean accuracy (%) of each module is presented in Table 5. The result indicates that the performance of the residual stream has greatly improved after adding the GFA block with global pooling. Furthermore, the fusion of both residual and dense streams shows superior performance to the individual streams. Specifically, the fusion of residual stream with GFA block (average pooling) and dense stream indicates a better performance of 91.24% (shown in a bold letter). The comparison of performance among GFA residual stream, SF-dense stream, and DeReFNet on all benchmark datasets is presented in Table 6. The tabulation results show the effectiveness of the DeReFNet among all benchmarked datasets.

4.2.3. *Qualitative analysis using Grad-CAM*

To better visualize the advantage of the proposed DeReFNet, gradient-weighted class activation mapping (Grad-CAM) [25] technique is used. This technique allows to investigate the input image to be categorized, and shows which parts /pixels of the image contribute more to deciding the final class decision of the model. Fig. 7 represents two examples of two different datasets using the Grad-CAM technique. The visualization result shows that, among other dense and residual blocks, the last dense and residual block of the DeReFNet concentrates on the silent areas of the hand gesture image. Furthermore, the DeReFNet captures more efficient information from the hand gesture image compared to both separate streams. As a result, this analysis shows that the DeReFNet consolidates high-class discriminating information from two distinct streams to represent the gesture class accurately.

4.2.4. *Quantitative analysis*

The quantitative performance of the proposed CNN is evaluated using the LOO CV test in terms of mean accuracy on all benchmarked datasets. The results are presented in Table 6. The result proves the effectiveness of the DeReFNet, which provides better generalization behaviour on a uniform and complex backgrounds based on vision-based hand gesture recognition datasets. The confusion matrices on the above results are shown in Fig. 8. In the JT-ASL dataset, only one gesture of class 'd' is miss-classified as class 'i'. This is because, during data acquisition, one subject used the index finger rather than the little finger to show the gesture pose 'd'. For the MU-ASL-digit dataset, the performance is restricted due to the class '4' and '5'. The gesture poses 'd' and 'h' are miss-classified in the NUS II dataset. In the MU-ASL-Alphabet dataset, the 34.3% of gesture class 'm' is miss-classified as 'n' class. The most confusing gestures for the ASL-FS colour dataset are 'k' and 'v'.

4.2.5. *Comparisons with other CNNs*

This section evaluates the compression performance of the DeReFNet with other CNNs on the MU-ASL-alphabet dataset. A comparison of number of trainable parameters, recognition accuracy (%), and average inference time (ms) between the DeReFNet and other fine-tuned CNNs such as VGG 16, ResNet 50, and DenseNet 201 are presented in the Table 7. The pre-trained CNNs are fine-tuned on training gesture samples of the MU-ASL-alphabet dataset keeping all the training hyper-parameters the same as defined by the proposed

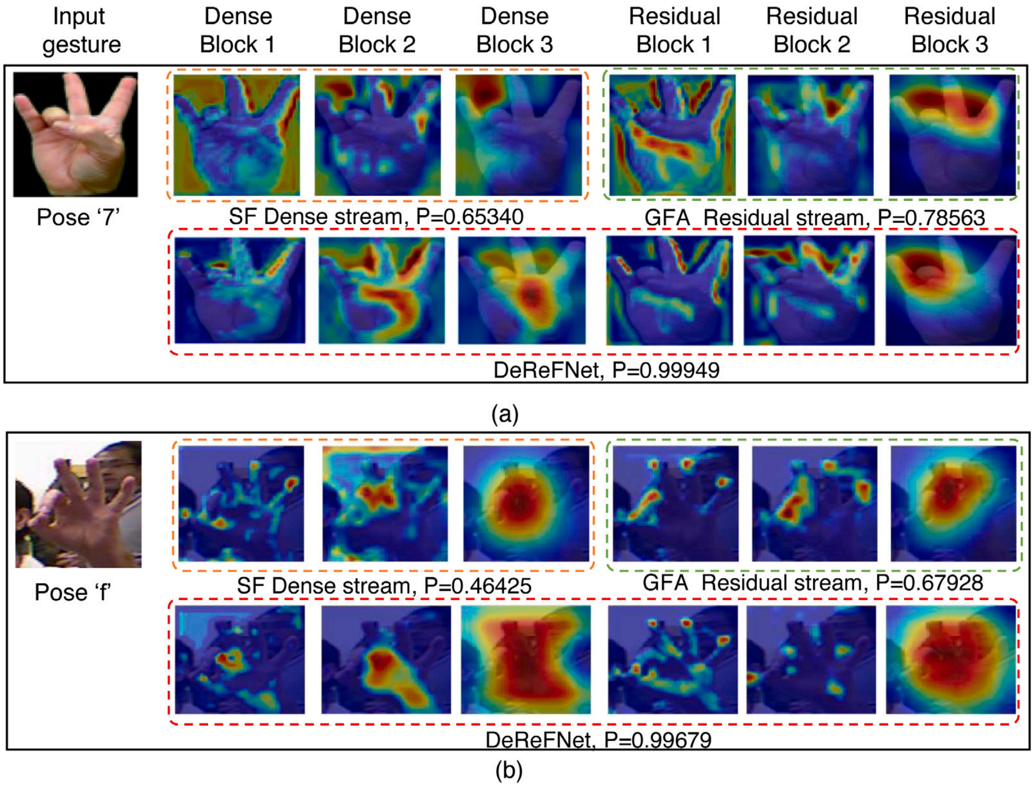

Fig. 7. Qualitative results of DeReFNet compared with SF dense stream and GFA residual stream using gradCAM. The visualization result is obtained for each block in the respective network architecture. The dense and residual block numbers are given in the top row of the image, and P denotes the softmax score of the network. Visual results of (a) MU-ASL-digit gesture pose '7', and (b) ASL-FS-colour gesture pose 'f'.

Table 6
Performance comparison among proposed DeReFNet with individual streams in terms of mean accuracy (%) on the benchmarked datasets.

Dataset	GFA-residual stream	SF-dense stream	Proposed DeReFNet
JT-ASL	98.75	97.92	99.58 ± 0.01
MU-ASL-digit	93.42	92.57	96.14 ± 1.12
MU-ASL-alphabet	93.88	92.39	95.70 ± 1.13
NUS-II	90.20	89.95	96.70 ± 1.65
ASL-FS-colour	88.45	86.95	91.24 ± 1.83

Table 7
Comparative analysis of proposed DeReFNet with state-of-the-art CNNs in terms of trainable parameters, mean accuracy, and average inference time on MU-ASL-alphabet dataset.

CNN models	Trainable parameters in millions	Mean accuracy (%)	Average inference time (ms)
VGG 16 [26]	138.35M	93.71	12.40
ResNet 50 [17]	25.58M	92.62	14.28
DenseNet 201 [18]	20.03M	91.86	51.34
SENet [27]	28.1M	93.87	21.36
Proposed DeReFNet	5.24M	95.70	9.54

DeReFNet. The above table shows that the proposed CNN architecture achieves 95.70% of mean accuracy with 5.24 M trainable parameters. The average inference time is the time taken by a technique to recognize a test gesture pose. The average inference time for fine-tuned VGG 16, ResNet 50, DenseNet 201 and DeReFNet are 12.40 ms, 14.28 ms, 51.34 ms and 9.54 ms respectively. The result justifies that the DeReFNet comprises less trainable parameters and achieves a faster recognition rate than other CNN architectures.

4.3. Performance analysis on similar gestures

In this work, the performance of DeReFNet is compared with the CNN using each stream of the network individually. That means the CNN with only GFA residual stream and the CNN with only SF dense stream. The comparison of inter-class similar gestures for various datasets is presented in Table 8. The result indicates that the DeReFNet can distinguish the inter-class similar gestures more accurately compared to the CNN architectures with the individual stream. This tabulation result proves the potentiality of the proposed DeReFNet.

4.4. Performance comparison of DeReFNet with occlusion on bench-marked datasets

In the field of static gesture recognition, it is always assumed and recommended for an occlusion-free environment. This is because the palm region is very small, and each part of the palm and fingers are extremely important to distinguish the gesture class. However, to analyse the exact impact of occlusion on the performance, a study is conducted. As most of the state-of-the-art datasets do not have occlusion, an artificial occlusion is injected into each of the image data as reported in [28,29]. In the experiment, two different occlusion patches of size 16 × 16 and 32 × 32 are used as shown in Fig. 9. The occluded gesture samples are then tested in the trained DeReFNet and the performance accuracy is presented in Table 9. It is clearly observed that the performance is largely affected due to the occlusion. However, a small occlusion, such as 16 × 16, can still provide satisfactory results.

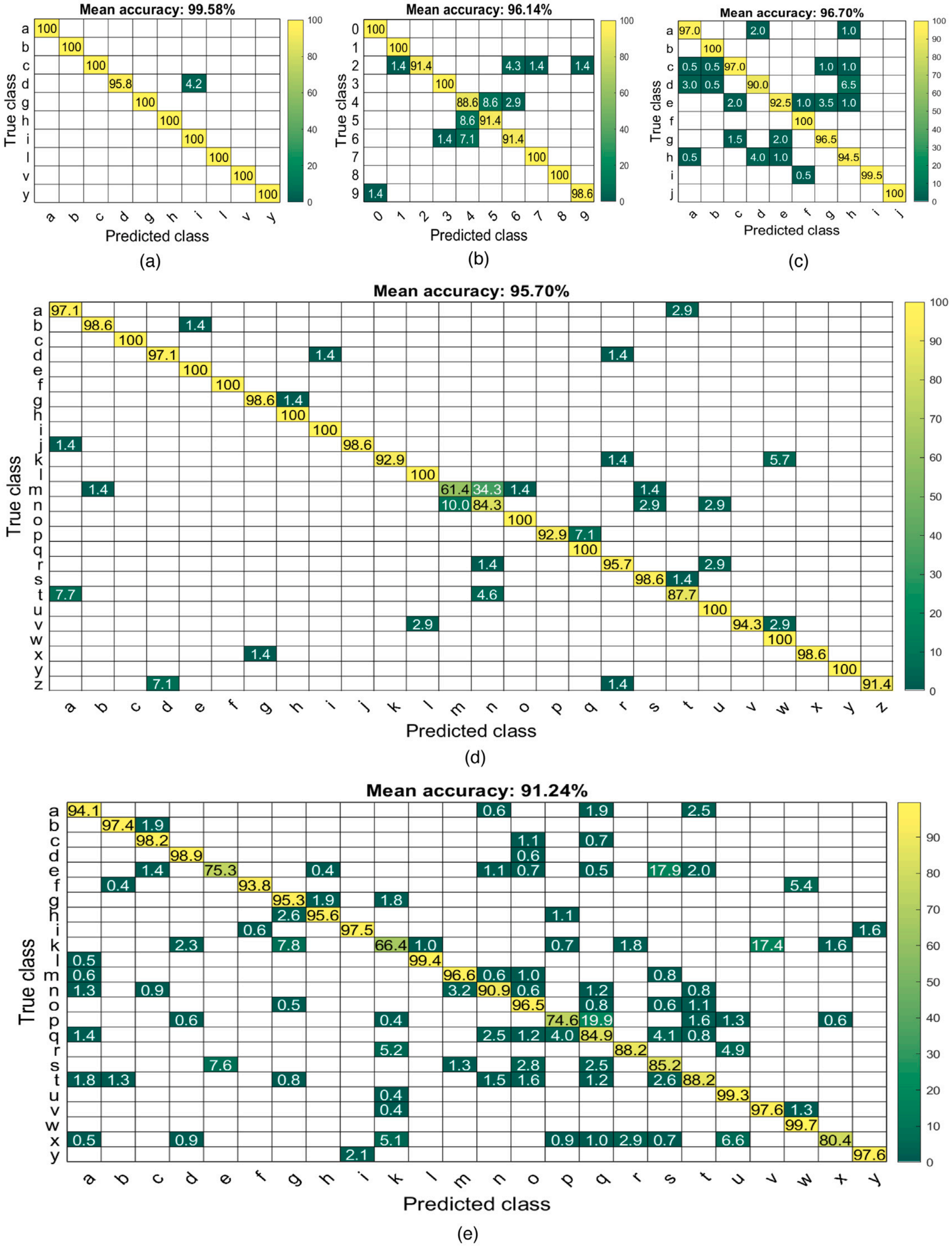

Fig. 8. Confusion matrices of the proposed DeReFNet on the benchmarked datasets. (a) JT-ASL, (b) MU-ASL-digit, (c) NUS-II, (d) MU-ASL-alphabet, and (e) ASL-FS-colour datasets.

Therefore, it is always recommended to avoid occlusion in case of vision based static hand gesture recognition.

4.5. Comparison with earlier techniques on bench-marked datasets

The performance comparison between the proposed approach with the state-of-the-art methods on standard datasets is presented in Table 10. In the JT-ASL dataset, the performance of the proposed DeReFNet is compared with the earlier reported techniques [3,8,30] using the LOO CV test. The results indicate that the mean accuracy (%) result of DeReFNet is 99.58% which is superior to the other reported techniques. The comparison between the existing methods [11,23,31] and the proposed technique on MU-ASL-digit dataset shows that the DeReFNet achieves superior performance to other existing methods. For the MU-ASL-alphabet dataset, the DeReFNet achieves 3.10% superior mean accuracy than the fusion of 'FC6' layer features from two pre-trained CNNs [22], 19.45% higher than the CNN [32] technique. For complex backgrounds dataset NUS-II, the DeReFNet outperform other state-of-the-art methods [16,20,22]. Similarly, ASL-FS-colour dataset achieves superior mean accuracy than other existing methods such as ExtriDeNet [16], and RBI-2RCNN [20]. The proposed DeReFNet perform superior to the other existing methods on all benchmark

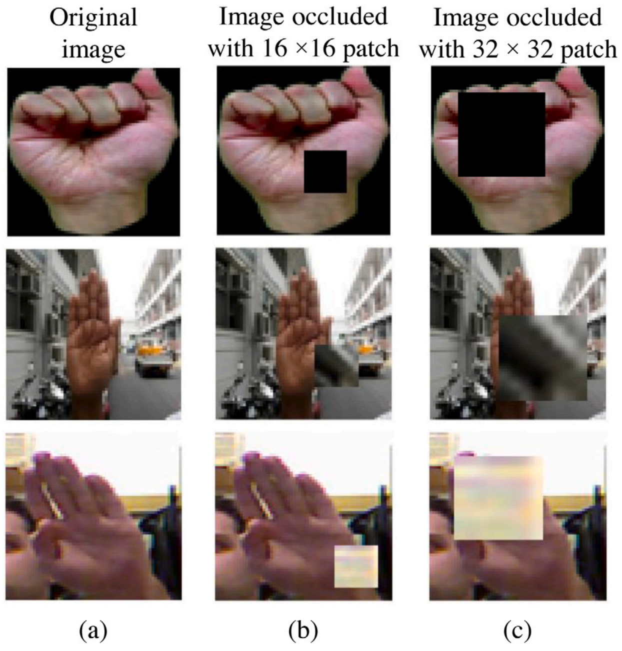
Original image
Image occluded with 16 ×16 patch
Image occluded with 32 × 32 patch
(a)
(b)
(c)

Fig. 9. Visualization of artificially created occluded hand gesture samples (a) Original, (b) 16 × 16 occlusion, and (c) 32 × 32 occlusion.

Table 8
Performance comparison (mean accuracy (%)) among the most similar gesture poses on the benchmarked datasets.

Dataset	Similar classes	GFA-residual stream	SF-dense stream	Proposed DeReFNet
JT-ASL	d	91.67	95.83	95.83
	i	100	100	100
	g	100	95.83	100
	h	95.83	91.67	100
MU-ASL-digit	7	97.14	92.86	100
	8	95.71	92.86	100
MU-ASL-alphabet	m	51.43	57.14	61.43
	n	62.86	64.29	84.29
NUS-II	c	87.50	91.00	97.00
	e	87.50	83.50	92.50
	d	93.00	93.00	90.00
	g	87.00	85.50	96.50
ASL-FS-colour	m	51.43	57.14	96.06
	n	62.86	64.29	90.90

Table 9
Performance comparison of DeReFNet with occlusions on the benchmarked datasets.

Dataset	DeReFNet without occlusion (%)	DeReFNet with 16 × 16 occlusion patch (%)	DeReFNet with 32 × 32 occlusion patch (%)
JT-ASL	99.58	92.91	47.50
MU-ASL-digit	96.14	91.29	51.28
MU-ASL-alphabet	95.70	90.75	43.13
NUS-II	96.70	89.65	31.45
ASL-FS-colour	91.24	84.07	49.01

datasets due to the following reasons: (i) The DeReFNet combines GFA-residual and dense stream, which explores the fusion of distinguished information from two different stream which represents the gesture image in a better way, (ii) The proposed GFA block in the residual stream gathers the low level and high-level information from each residual block for better HGR.

Table 10
Performance comparison (mean accuracy (%)) of proposed technique with state-of-the-art techniques using LOO CV test on benchmarked datasets.

Dataset	Methods	Mean accuracy (%)
JT-ASL	DWT F-ratio and SVM [3]	95.42
	Elastic Graph Matching [8]	94.3
	LBP and MLP [30]	93.3
	DCT, DWT, PCA, and PNN [33]	94.00
	AlexNet ’FC6’ + VGG-16 ‘FC6’ and SVM [22]	89.17
	Proposed DeReFNet	**99.58**
MU-ASL-digit	NMF and CS [31]	87.80
	SoGF-FV and SVM [11]	95.30
	AlexNet ‘FC6’ + PCA and SVM [23]	95.00
	Proposed DeReFNet	**96.14**
MU-ASL-alphabet	AlexNet ‘FC6’ + PCA and SVM [23]	92.60
	Hybrid DWT-Gabor filter and KNN [32]	50.83
	CNN [32]	76.25
	Proposed DeReFNet	**95.70**
NUS-II	AlexNet ’FC6’ + VGG-16 ‘FC6’ and SVM [22]	92.65
	ExtriDeNet [16]	61.49
	RBI-2RCNN [20]	94.80
	Proposed DeReFNet	**96.70**
ASL-FS-colour	ExtriDeNet [16]	81.72
	RBI-2RCNN [20]	88.65
	Proposed DeReFNet	**91.24**

5. Conclusion

This paper proposes a dual-stream dense residual fusion network known as DeReFNet for the accurate recognition of hand gestures in a vision-based environment. The DeReFNet combines the global information from the receptive fields of each residual block in one stream and dense spatial information from another. The combination of both streams explores the accurate finger information in the gesture images. Furthermore, the developed architecture can reduce the inter-class variations in the hand gestures observed visually using the GradCAM technique and form experimental results. In addition, the developed architecture has less trainable parameters and requires less average inference time than other state-of-the-art CNNs. Various parameter studies, ablation studies, and qualitative and quantitative analysis demonstrate that the proposed DeReFNet outperforms the earlier reported techniques on four publicly available benchmarked datasets. Although the proposed technique achieves promising performance and handles most of the closely related gestures, the work of this paper tries to motivate more future works to handle the misclassified gestures.

CRediT authorship contribution statement

Jaya Prakash Sahoo: Conceptualization, Investigation, Methodology, Writing – original draft. **Suraj Prakash Sahoo:** Review & editing. **Samit Ari:** Resources, Supervision, Review & editing. **Sarat Kumar Patra:** Supervision, Review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

References

[1] P.K. Pisharady, M. Saerbeck, Recent methods and databases in vision-based hand gesture recognition: A review, Comput. Vis. Image Underst. 141 (2015) 152–165, http://dx.doi.org/10.1016/j.cviu.2015.08.004.

[2] E. Ohn-Bar, M.M. Trivedi, Hand gesture recognition in real time for automotive interfaces: A multimodal vision-based approach and evaluations, IEEE Trans. Intell. Transp. Syst. 15 (6) (2014) 2368–2377.

[3] J.P. Sahoo, S. Ari, D.K. Ghosh, Hand gesture recognition using DWT and F-ratio based feature descriptor, IET Image Process. 12 (10) (2018) 1780–1787, http://dx.doi.org/10.1049/iet-ipr.2017.1312.

[4] C. Wang, Z. Liu, S.-C. Chan, Superpixel-based hand gesture recognition with kinect depth camera, IEEE Trans. Multimed. 17 (1) (2015) 29–39, http://dx.doi.org/10.1109/tmm.2014.2374357.

[5] L. Guo, Z. Lu, L. Yao, Human-machine interaction sensing technology based on hand gesture recognition: A review, IEEE Trans. Hum.-Mach. Syst. 51 (4) (2021) 300–309, http://dx.doi.org/10.1109/THMS.2021.3086003.

[6] D.A. Reddy, J.P. Sahoo, S. Ari, Hand gesture recognition using local histogram feature descriptor, in: 2018 2nd International Conference on Trends in Electronics and Informatics, ICOEI, IEEE, 2018, pp. 199–203.

[7] D.K. Ghosh, S. Ari, On an algorithm for vision-based hand gesture recognition, Signal, Image and Video Process. 10 (4) (2016) 655–662.

[8] J. Triesch, C. von der Malsburg, Classification of hand postures against complex backgrounds using elastic graph matching, Image Vis. Comput. 20 (13) (2002) 937–943.

[9] A. Barczak, N. Reyes, M. Abastillas, A. Piccio, T. Susnjak, A new 2D static hand gesture colour image dataset for ASL gestures, Res. Lett. Inf. Mathematical Sci. 15 (2011) 12–20.

[10] P.K. Pisharady, P. Vadakkepat, A.P. Loh, Attention based detection and recognition of hand postures against complex backgrounds, Int. J. Comput. Vis. 101 (3) (2013) 403–419.

[11] L. Fang, N. Liang, W. Kang, Z. Wang, D.D. Feng, Real-time hand posture recognition using hand geometric features and fisher vector, Signal Process., Image Commun. 82 (2020) 115729, http://dx.doi.org/10.1016/j.image.2019.115729.

[12] M. Lazarou, B. Li, T. Stathaki, A novel shape matching descriptor for real-time static hand gesture recognition, Comput. Vis. Image Underst. 210 (2021) 103241, http://dx.doi.org/10.1016/j.cviu.2021.103241.

[13] Z. Wang, Gesture recognition by model matching of slope difference distribution features, Measurement 181 (2021) 109590, http://dx.doi.org/10.1016/j.measurement.2021.109590.

[14] Y.S. Tan, K.M. Lim, C.P. Lee, Hand gesture recognition via enhanced densely connected convolutional neural network, Expert Syst. Appl. 175 (2021) 114797, http://dx.doi.org/10.1016/j.eswa.2021.114797.

[15] W. Zhou, K. Chen, A lightweight hand gesture recognition in complex backgrounds, Displays 74 (2022) 102226.

[16] G. Bhaumik, M. Verma, M.C. Govil, S.K. Vipparthi, ExtriDeNet: an intensive feature extrication deep network for hand gesture recognition, Vis. Comput. (2021) 1–14, http://dx.doi.org/10.1007/s00371-021-02225-z.

[17] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., 2016, pp. 770–778.

[18] G. Huang, Z. Liu, L. Van Der Maaten, K.Q. Weinberger, Densely connected convolutional networks, in: Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., 2017, pp. 4700–4708.

[19] G. Philipp, D. Song, J.G. Carbonell, Gradients explode-deep networks are shallow-resnet explained, in: Proc. 6th Int. Conf. Represent ICLR Workshop Track, 2018.

[20] J.P. Sahoo, S.P. Sahoo, S. Ari, S.K. Patra, RBI-2RCNN: Residual block intensity feature using a two-stage residual convolutional neural network for static hand gesture recognition, Signal, Image and Video Process. 16 (8) (2022) 2019–2027, http://dx.doi.org/10.1007/s11760-022-02163-w.

[21] S.F. Chevtchenko, R.F. Vale, V. Macario, F.R. Cordeiro, A convolutional neural network with feature fusion for real-time hand posture recognition, Appl. Soft Comput. 73 (2018) 748–766, http://dx.doi.org/10.1016/j.asoc.2018.09.010.

[22] J.P. Sahoo, S. Ari, S.K. Patra, A user independent hand gesture recognition system using deep cnn feature fusion and machine learning technique, in: New Paradigms in Computational Modeling and Its Applications, Elsevier, 2021, pp. 189–207, http://dx.doi.org/10.1016/B978-0-12-822133-4.00011-6.

[23] J.P. Sahoo, S. Ari, S.K. Patra, Hand gesture recognition using PCA based deep CNN reduced features and SVM classifier, in: 2019 IEEE International Symposium on Smart Electronic Systems (ISES)(Formerly INiS), IEEE, 2019, pp. 221–224, http://dx.doi.org/10.1109/iSES47678.2019.00056.

[24] N. Pugeault, R. Bowden, Spelling it out: Real-time ASL fingerspelling recognition, in: 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops), IEEE, 2011, pp. 1114–1119, http://dx.doi.org/10.1109/ICCVW.2011.6130290.

[25] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 618–626.

[26] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, 2014, arXiv preprint arXiv:1409.1556.

[27] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7132–7141.

[28] J. Zhuo, Z. Chen, J. Lai, G. Wang, Occluded person re-identification, in: 2018 IEEE International Conference on Multimedia and Expo, ICME, 2018, pp. 1–6, http://dx.doi.org/10.1109/ICME.2018.8486568.

[29] S.P. Sahoo, S. Modalavalasa, S. Ari, DISNet: A sequential learning framework to handle occlusion in human action recognition with video acquisition sensors, Digit. Signal Process. 131 (2022) 103763, http://dx.doi.org/10.1016/j.dsp.2022.103763.

[30] K. Sadeddine, R. Djeradi, F.Z. Chelali, A. Djeradi, Recognition of static hand gesture, in: 6th International Conference on Multimedia Computing and Systems, ICMCS, 2018, pp. 1–6, http://dx.doi.org/10.1109/ICMCS.2018.8525908.

[31] H. Zhuang, M. Yang, Z. Cui, Q. Zheng, A method for static hand gesture recognition based on non-negative matrix factorization and compressive sensing, IAENG Int. J. Comput. Sci. 44 (1) (2017) 52–59.

[32] V. Ranga, N. Yadav, P. Garg, American sign language fingerspelling using hybrid discrete wavelet transform-gabor filter and convolutional neural network, J. Eng. Sci. Technol. 13 (9) (2018) 2655–2669.

[33] K. Sadeddine, F.Z. Chelali, R. Djeradi, Sign language recognition using PCA, wavelet and neural network, in: 2015 3rd International Conference on Control, Engineering & Information Technology, CEIT, IEEE, 2015, pp. 1–6.