International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Traitement du Signal
Vol. 40, No. 2, April, 2023, pp. 647-655
Journal homepage: http://iieta.org/journals/ts

Human Face and Facial Expression Recognition Using Deep Learning and SNet Architecture Integrated with BottleNeck Attention Module

Sumithra Meenatchi Sundaram*, Rajkumar Narayanan

Department of Computer Science and Engineering, Vel Tech Rangarajan Dr. Sagunthala Institute of Science and Technology, Avadi, Chennai 600062, Tamilnadu, India

Corresponding Author Email: vtd702@veltech.edu.in

https://doi.org/10.18280/ts.400223

Received: 24 November 2022
Accepted: 18 March 2023

Keywords:
facial expression recognition, face recognition, SNet architecture, BottleNeck attention module, CNN, thermal images

ABSTRACT

Thermal infrared face image recognition with the help of deep learning technology has become the most debated concept in research area nowadays. Many articles are done and being working on this area to discover novel findings. Thermal infrared images can be recognised irrespective of light conditions, aging and facial disguises. This paper proposes a method named SNet integrated with BottleNeck Attention Module (SN-BNAM) for thermal face image recognition using SENet architecture in which the BottleNeck Attention Module is integrated. After squeeze and excitation process, the channel and spatial attention is inferred as two separate branches inside the BottleNeck Attention Module (BAM). This module is placed at each BottleNeck area. The SN-BNAM module can be integrated with any feed forward convolutional neural networks. The efficiency of the proposed system is evaluated by experimenting on various architectures and object validation is done on VOC 2007, MS COCO, CIFAR-100 and ImageNet-1K datasets. These experiments proves that our method shows consistent improvement in image classification and object detection.

1. INTRODUCTION

Human face recognition is a biometric application, which is being used to access various applications such as unlocking mobile phones, attendance, healthcare and security systems. Many traditional methods have been used in earlier days for authentication and security which faces many challenges. For example, passwords and PINs are hard to remember they can be easily speculated stolen or forgotten. While smart cards, tokens and plastic cards may be misplaced or reproduced, magnetic cards get tarnished or illegible. An efficient technique is necessary to overcome the ambiguities that exists in traditional method of authentication. To overcome these difficulties biological characteristics and traits of human are used for authentication such as fingerprint, iris and face recognition. Among these choices face recognition is more significant and promising biometric feature for authentication and security. The wide use of smart phones and digital cameras makes the face recognition process easier and more efficient. Face recognition enables an accurate, secure and fast authentication for security access and surveillance.

Many researchers have been done in visible spectrum (RGB) and infrared spectrum but the foresaid methods of face recognition are affected due to factors like illumination, light, darkness and occlusion [1]. Therefore, in order to overcome the above problems, thermal images are used for face recognition, which works well in any lightning conditions. The thermal images are captured between the wavelengths 8µm to 12µm [2]. Human thermal faces images are created by the heat patterns that are emitted from the body and these images are autonomous of the environmental lighting conditions [3].

Artificial Intelligence (AI) refers to a human like behaviour portrayed by a system or a machine. Programming the computers to exhibit the mimic of human behaviour is the main task in AI, which is achieved using massive data from the previous samples of identical behaviour. AI is excelled by various algorithms using Machine Learning (ML) and Deep Learning (DL). Machine Learning is a subdivision of AI, which automates the execution of analytical model building and prepares machines to adapt the new model independently. Deep Learning on the other hand, is a subset of ML that performs the tasks more superior than the traditional machine learning approaches. DL mainly composites two key techniques: supervised or unsupervised learning [4]. DL implements a of multiple hidden layer artificial neural networks, which takes the output of the previous layer as the input of the next layer. DL is viewed as a tool to enhance the results and optimization of processing time in various computational process. Nowadays, Deep Learning (DL) is emerging as an effective tool for face recognition and shows impressive results. The Convolutional Neural Network (CNN) is a type of deep neural networks, which extract visual feature automatically [5] DL has made the automation process of choosing the filters to extract best features from the image and gives better accuracy. Applications of DL include, Image Recognition, Natural Language Processing, Recommendation systems and speech recognition [6]. This paper focus on recognition of thermal face images of human using Deep Learning (DL) embedded with SNet architecture. The SNet architecture consists of a special block called squeeze and excitation (SE) block that can increase the representational ability of any network through explicit modelling of the interdependencies between channels. An SE network is constructed by forming a stack of multiple SE blocks. The function of the SE block is to compute the channel attention and gives an increasing performance gain with low cost.

SENet provides an effective way of channel attention computation along with performance excellence. The main purpose of SENet is to perform the cross-channel relationships by learning the modulation weights in each channel [7]. While SNet architecture works well for Human Face recognition, the proposed method SENet-BottleNeck Attention Module (SN-BNAM) increases the accuracy and reduces the error rates. A BottleNeck Attention Module (BNAM), is placed between each SE block, where the information flow is critical. Before getting in to BNAM the concept of attention module has to be explained. Attention modules or attention mechanism are DL techniques that gives an added focus in a particular component which has specific importance. An attention module is used to elevate the performance of Convolutional Neural Networks (CNN) by focusing it only on the features that are more important and eliminating the unnecessary features [8]. Generally, attention mechanism is applied to channel and spatial dimensions. The proposed SN-BNAM method adopts the BottleNeck attention module by Park et al. [9]. A BottleNeck attention module can be adapted with any CNNs. In this proposed system, the BAM is induced with SENet architecture to achieve performance accuracy. The BAM divides the given 3d feature map into two attention modules, called the channel attention and spatial attention module, which can be considered as the feature detectors to gain the explicit knowledge of where and what to be focused. Experiments are done with our own Thermal Human Face Data Set (THFDS) dataset on various baseline architecture and the results shows that by adding SN-BNAM with the baseline elevates the efficiency. The performance of the proposed method is also evaluated by implementing it on CIFAR-100 and ImageNet-1K datasets and the results are reported. Finally, the performance improvement of object detection task is done on VOC 2007 and MS COCO datasets indicates the excellence of SN-BNAM.

1.1 Problem statement

In the area of human face recognition systems, many researches are done using various algorithms and architectures. Massive researches in face recognition for RGB images are done and achieved practical success too. These studies have the challenges of illusion, occlusion, light changes and disguises. The studies conducted by previous researches used multiple face images that have varying poses and lighting conditions to achieve maximum accuracy. This process requires a huge database with multiple images of the same person causing greater storage requirements and complex computations. Computations on such huge databases are error prone, time consuming and reduces the performance accuracy. While face recognition using thermal images are carried out in many studies, the proposed methods not only recognize face images and recognize the basic facial expressions under any circumstances. Many traditional algorithms have been used for face recognition, but they need high computational power, recognition process may include unnecessary features, reduced accuracy and higher error rates.

To overcome these problems the proposed method uses thermal infrared images for face recognition.

1.2 Contributions

The main contributions of this paper are as follows:

(1) This article uses thermal infrared images for face recognition to overcome the problems such as, illusion, pose variation, occlusion, expressions and aging.

(2) The proposed method SN-BNAM implements the SNet architecture, which integrates the lightweight BottleNeck attention module.

(3) In the proposed method, the SE block enhances the interdependencies of the channel with mere computational cost and boosts the performance of the recognition process.

(4) The lightweight BottleNeck attention module focuses on incorporating channel attention with spatial attention to achieve performance gains.

(5) The proposed system also detects the basic face expressions happy, sad, natural, fear and disguise.

(6) Our proposed method works well under various environmental conditions such as indoor, outdoor, day and night.

The rest of the article is organised in the following way. Section 2 lists out the literature review, section 3 displays the methodology of the proposed system, section 4 explains the results of this proposed system and section 5 gives the conclusion.

2. LITERATURE REVIEW

The performance of infrared face recognition systems is impacted by the temporal variations present in thermal face images, which are primarily caused by various environmental factors, physiological changes in the subjects, and variations in the responsiveness of the infrared detectors at the time of the capture. These five techniques local binary pattern (LBP), Weber linear descriptor (WLD), Gabor jet descriptors, scale invariant feature transform, and sped-up robust features have been used to develop thermal face recognition systems [1]. Recent advancements in the technology for deep learning in the field of thermal infrared face recognition have made it possible for the numerous research groups working on this topic to produce numerous cutting-edge results [2]. In addition to being able to identify faces that cannot be seen in visible light, thermal infrared face recognition can also identify the facial blood vessel structure. It reviews earlier studies on temperature variations, mathematical equations, wave types, and techniques for thermal infrared face recognition [3]. A new area of machine learning (ML) research is deep learning. In large databases, the deep learning methodology applies nonlinear transformations and high-level model abstractions. Deep learning architectures have recently made significant strides in a variety of fields, and these developments have already had a big impact on artificial intelligence [4]. Face recognition using convolutional neural networks (CNNs) for daily attendance taking, however, continues to be a difficult problem due to the challenges of sample collection. The samples can be increased through data augmentation, which has been used in small sample learning [5]. The intelligent process of identifying and authenticating people with occluded faces has been greatly aided by advancements in machine learning in MFR. The characteristics of deep network architectures and deep feature extraction strategies are used to introduce cutting-edge techniques [6]. Triplet attention encodes inter-channel and spatial information with minimal computational overhead for an input tensor while also creating inter-dimensional dependencies through rotation operations and residual transformations [7]. Convolutional neural network with a Convolutional block attention module (CBAM) for finger vein recognition, which can achieve a more accurate capture of visual structures through an attention mechanism and account for the different importance of pixels [8]. The BottleNeck Attention Module (BAM) that can be integrated with any feed-forward Convolutional neural network is a straightforward and efficient attention module. The broad applicability of BAM by consistently improving classification and detection performances with different models [9]. Squeeze-and-Excitation Network (SENet) architectures that generalise incredibly well across difficult datasets. Most importantly, we discover that SE blocks significantly improve performance for current state-of-the-art deep architectures at a low additional computational cost [10]. Convolutional neural networks by creating an AW-convolution, where the shape of attention maps matches that of the weights rather than the activations, in order to jointly solve the two problems. A complementary approach to earlier attention-based strategies, like those that use the attention mechanism to investigate the connection between channel-wise and spatial features, is our proposed attention module [11]. Convolutional Block Attention Module (CBAM) is the attention module that we suggest being used with all feed-forward convolutional neural networks. CBAM can be seamlessly and with minimal over heads integrated into any CNN architecture because it is a lightweight and general module [12]. Residual networks are simpler to optimise and can improve accuracy through significantly more depth. On the ImageNet dataset, we test residual nets up to 152 layers deep, which is eight times deeper than VGG nets while still being less complex [13]. ResNeXt further using the COCO detection set and the ImageNet-5K set, both of which exhibit superior performance to ResNet [14]. As a family of very deep architectures with impressive accuracy and nice convergence behaviours, deep residual networks have emerged. Using identity mappings as the skip connections and after-addition activation, we analyse the propagation formulations underlying the residual building blocks and find that the forward and backward signals can be transmitted directly from one block to any other block [15]. The overall summary of the literature review is listed in Table 1.

Table 1. Summary of literature review

S.No	Year	Problem	Solution / Review
1.	2017	Thermal Face Recognition Under Temporal Variation Conditions	Comparison of five methods for FR is done. Weber Linear Descriptor performs better.
2	2021	Thermal Infrared Face Recognition	Review: Thermal IR face recognition is more precise than visible light and much developmental research has to be done.
3.	2022	A survey on thermal face recognition using machine learning	Study of face recognition with visible and thermal infrared images. CNN approach for thermal face images gives better result.
4.	2019	Face Recognition via Deep Learning	Solution: Data Augmentation Based on Orthogonal Experiments. Limitation: Accuracy % is less than 90.
5.	2018	Face recognition using deep learning methods a review	Review: Deep Learning models perform better with larger dataset.
6.	2022	Finger Vein Recognition	CNN with Convolutional Block attention Module. Limitation: Done only with minimum Data.
7.	2021	Residual attention network for image classification	Trunk-and-mask encoder-decoder module. Limitation: Computationally complex due to the direct generalisation of 3D attention maps.
8.	2019	Gcnet: Non-local networks meet squeeze-excitation networks and beyond	Solution: Novel NL block integrated with SE block. Limitation: Uses complex permutation based operations to reduce the feature maps.
9.	2021	Convolutional Triplet Attention Module	Solution: Triplet attention with 3 branches. Limitation: Complex network structure.
10.	2017	Densely Connected Convolutional Networks	Solution: DenseNet- collects all layers directly with each other. Limitation: Focus is on depth, width and cardinality and attention fact not considered.
11.	2017	Image recognition	Solution: Squeeze and Excitation block. Limitation: Spatial attention which decides “where” to focus is missing.

3. METHODOLOGY

This article proposes the SNet architecture integrated with BottleNeck Attention module (SN-BNAM) method for human face recognition. The proposed method uses single human face for face recognition and for the data set, our own dataset Thermal Human Face Dataset (THFDS) is used, which consists of thermal human face images. Thermal images overcome the challenges of other spectrums such as illusion, facial disguises, aging etc. This article also focusses on face recognition under different environmental conditions such as indoor, outdoor, day and night. In addition to face recognition, facial expression recognition is also executed considering the basic expressions such as happy, sad, natural, fear and disguise. Experiments are carried out considering all the above factors and the results are reported in terms of accuracy, error rates, intensity, precision, recall value and F1-score. Figure 1 displays the block diagram of the proposed method.

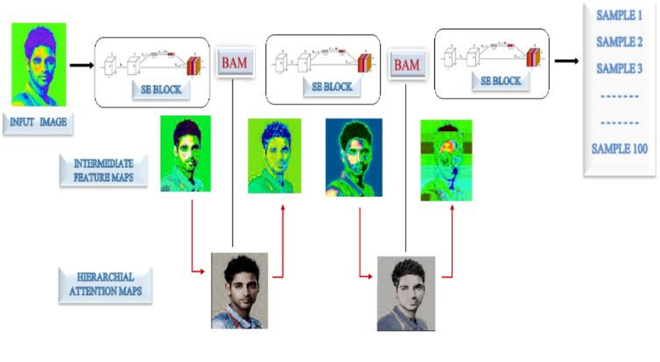


Figure 1. Proposed SN-BNAM method

3.1 The SENet architecture

Figure 2 depicts a model of the squeeze and excitation block. From the figure, it can be observed that any transformation Ftr, that maps the input X to U (feature map) where U∈ R H×W×C, it is necessary to execute a feature recalibration, which is performed in the SE block [10].

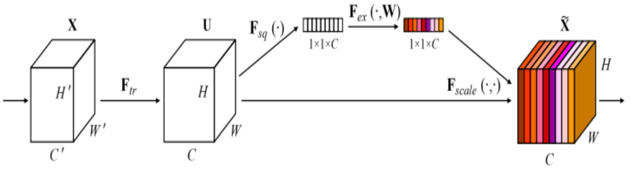


Figure 2. The squeeze and excitation block of SNet [10]

The input X is mapped with the feature map U, in which U $\in \mathbb{R}^{H\times W\times C}$. Initially the squeeze operation is applied on the features U which results a channel descriptor through the aggregation of feature maps along the spatial dimensions (H x W). This descriptor performs the function of producing the integration of global distribution of channel wise responses and allows all the layers in the network to use the information of the global recipient field in the network. This aggregation is continued by the excitation operation, which aids in the form of a basic self-gating tool that accepts the embedding as its input and yields a cluster of per channel modulation weights. Now, these weights are employed to the feature map U to get the output of the squeeze and excitation block, which can be given directly to the consequent layers of the network.

3.1.1 Working of the Squeeze and Excitation Block (SEB)

A SEB forms a computational unit that is set upon a transformation F_{tr} mapping to an input X to the feature maps U $\in \mathbb{R}^{H\times W\times C}$. In the following equation F_{tr} is taken as the convolutional operator and the set of filter kernels is denoted by V=[v_1, v_2,...,v_C]. The output can be written as U=[u_1, u_2,...,u_C] in which [10]:

$$\boldsymbol{u_c} = \mathbf{v}_c * \mathbf{X} = \sum_{s=1}^{C'} \mathbf{v}_c^s \mathbf{x}^s \qquad (1)$$

In the above equation the * operation denotes convolution, v s c is a two dimensional spatial kernel that represents a single channel and acts on the respective channel X and $v_c=[v_c^1, v_c^2$ c,...,$v_c^{C'}]$, X=[x_1, x_2,...,$x^{C'}$] and $u_c \in \mathbb{R}^{HxW}$. The output is generated by the implementation of the summation of all the channels and hence the channel dependencies are embedded in v_c implicitly. The convolution model depicts the channel relationship as an implicit and local one but it will be better if these convolutional features are explicit and the network has the capability to elevate its sensitive features. It is necessary to provide an explicit modelling to handle the issues of informative features being exploited by subsequent transformations. The SE block provides this by granting access to global information and the filter responses are recalibrated in two processes squeeze an excitation before feeding into the subsequent transformations.

3.1.2 The squeeze operation

To overcome the issue of the channel dependencies getting exploited, all learned filters operate with a local receptive field and therefore every single unit of the transformation output U will not be able to exploit the channel dependencies out of this region. This entire process is done by the squeeze operation, where the global spatial information is squeezed into a channel descriptor. A global average pooling is used to generate the statistics of each channel. Statistic $z \in \mathbb{R}^C$ is obtained by reducing U along its spatial dimensions HxW, where the c^{th} element of Z is computed by [10]:

$$z_c = \mathbf{F}_{sq}(\mathbf{u}_c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} u_c(i, j) \qquad (2)$$

The output U is viewed as a cluster of local descriptors and their statistics represents the entire image.

3.1.3 The excitation operation

The information collected by the squeeze operation is used in the next operation where the channel wise dependencies are fully captured. This objective is done with the following criteria: (1) the function should have the capability of learning nonlinear interaction between the channels, (2) a nonlinear mutually exclusive relationship must be learned. These criteria are met by employing a basic gating mechanism as follows [10]:

$$\boldsymbol{s} = \boldsymbol{F_{ex}}(\boldsymbol{z}, \boldsymbol{W}) = \mho(\boldsymbol{y}(\boldsymbol{z}, \boldsymbol{W})) = \mho(\boldsymbol{W_2}\boldsymbol{\delta}(\boldsymbol{W_1 z})) \qquad (3)$$

In the above sigmoid activation δ is the ReLU function, W1∈ ℝC/r X C and W2∈ℝ C× C/r. The gating mechanism is parameterised by a BottleNeck of two fully connected layers namely, the dimensionality reduction layer with the ration r and a dimensionality increasing layer returning back to the channel dimension of the output U. The block's final output is attained by rescaling U with the activation [10]:

$$\widetilde{X_c} = F_{scale}(u_c, s_c) = s_c u_c \qquad (4)$$

F_{scale} (u_c, s_c) refers to channel-wise multiplication between the scalar s_c and the feature map $u_c\mathbb{R}^{HxW}$.

3.2 BottleNeck attention module

Attention Module: An attention mechanism technique captures long-range of feature interactions and boosts the representations of the CNNs [11]. For any input image, the two attention modules, channel and spatial, compute the complementary attention to focus on what and where respectively. These two modules can be fixed in parallel or sequential to each other. Experimental results prove that keeping the channel attention module at first in sequential manner gives better results [12].

The attention modules are fixed, at the point of BottleNeck where down-sampling of feature maps happens. BAM constructs a hierarchical attention where BottleNeck occurs and it is trainable with feed forward models. In the proposed SN-BNAM method, this module is integrated with the SNet architecture and therefore the merits of both SNet and BAM are utilized for the face recognition process to achieve higher accuracy.

3.2.1 Channel attention

This module performs the function of feature extraction and reduces the data loss by squeezing feature maps. This process is done using the global average-pooling layer, which yield the overall information of the feature map **F** and produce the channel vector $F_C \in \mathbb{R}^{C\times1\times1}$. Using the channel vector $F_{C,}$ the attention is estimated using a multi-layer perception (MLP) with single hidden layer and the channel attention map $M_c(F) \in \mathbb{R}^{C\times1\times1}$ is produced. The parameter overhead is reduced by setting the hidden activation size is set to $\mathbb{R}^{C\times1\times1}$, where r is the reduction ratio. After applying the shared network, a batch normalization (BN) layer is added to make the scale adjusted with the spatial output. The channel attention is calculated as follows:

$$\mathbf{M_c(F)}\text{=BN(MLP(AvgPool(}\mathbf{F}\text{)))=BN(W1(w0AvgPool(F)+b0)+b1)} \quad (5)$$

In the above equation, $W_0 \in \mathbb{R}^{C/r\times C}$, $W_1 \in \mathbb{R}^{C\times C/r,}$ $b_0 \in \mathbb{R}^{C/r}$, $b_1 \in \mathbb{R}^{C}$. As the MLP network is a shared one the weights W_0 and W_1 were shared by both the inputs.

3.2.2 Spatial attention

The spatial attention module gives a spatial attention map $M_S(F) \in \mathbb{R}^{HxW}$ to suppress the features in various spatial locations. This module is used to find out which spatial locations are to be focused so that the important contextual information is preserved. There is a need for a massive receptive field that can leverage the contextual information efficiently. A dilated convolution is implemented to increase receptive field with greater efficiency. Dilated convolution constructs an effective spatial map compared to the standard convolution. The BottleNeck structure is adopted in the spatial branch which reduces the number of parameters and computational overhead. The feature F belongs to $\mathbb{R}^{CxHxW}$ is reduced to a dimension of $\mathbb{R}^{C/rxHxW}$ using the 1×1 convolution for integration and it compress the feature map across the channel dimension. Once the reduction process is completed (reduction ratio r), two 3×3 dilated convolutions are applied to make use of the information efficiently. Once again, the features are reduced to r1xhxw in the spatial attention map by using 1×1 convolution. A batch normalisation is applied at the end of the spatial branch for scale adjustment. The spatial attention is calculated as:

$$M_s(F) = BN\ (f_3^{1X1}\left(f_2^{3X3}\left(f_1^{3X3}\left(f_0^{1X1}(F)\right)\right)\right)) \quad (6)$$

where, f indicates the convolution operation, BM- the batch normalization operation and the superscripts of f denotes the size of the convolution filters. The two 1×1 convolutions are for channel reduction and the intermediate 3×3 dilated convolutions are applied to accumulate the contextual information in the massive receptive field.

3.2.3 Working of BAM

Figure 3 depicts the entire structure of BottleNeck Attention Module. Given the input feature map $F \in \mathbb{R}^{C\times H\times W}$, it infers a 3Dimensional attention map $M(F) \in \mathbb{R}^{C\times H\times W.}$ To develop an efficient module, first the channel attention $M_C(F) \in \mathbb{R}^{C}$ and spatial attention $M_S(F) \in \mathbb{R}^{H\times W}$ as two separate branches and finally compute M(F), the attention map as:

$$\text{M(F) as: M(F)=}\sigma\text{(Mc(F)+Ms(F))} \quad (7)$$

Once the channel attention and the spatial attention are acquired both are merged to yield the final 3Dimensional attention map **M(F)**. These two attention maps exhibit different shapes, therefore before merging they are expanded to $\mathbb{R}^{C\times H\times W}$. This merging is done using the element-wise summation and after that a sigmoid function is taken to get the final 3Dimensional attention map **M(F)** between the range 0 to 1.

The channel attention Mc and the spatial attention Ms are combined to compute the attention map M(F) by the equation:

$$\text{M(F) as: M(F)=}\sigma\text{(Mc(F)+Ms(F))} \quad (8)$$

The obtained 3D attention map **M (F)** $\mathbb{R}^{C\times H\times W}$ and the input feature map **F** are multiplied element-wisely and added to the original input feature map to obtain the refined feature map using the following equation where ⊗ indicates the element-wise multiplication.

$$F' = F + F \otimes M(F) \quad (9)$$

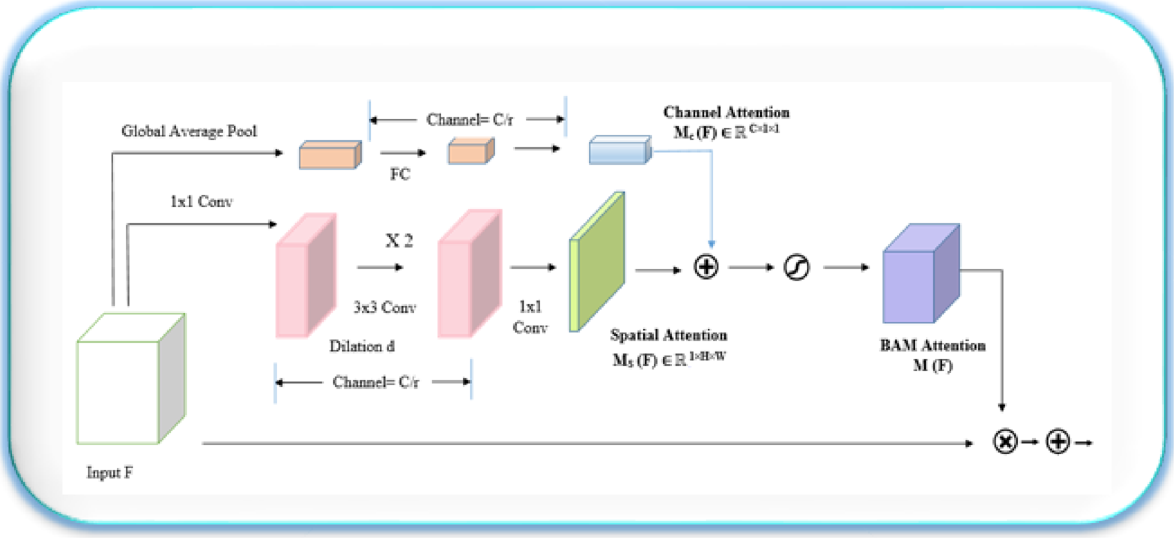


Figure 3. Detailed view of the BottleNeck attention module

4. RESULTS AND DISCUSSIONS

To evaluate the proposed SN-BNAM method, other network models such as ResNet18, ResNet50 are implemented with our data set THFDS and the experimental results are depicted in Table 2. As depicted in Table 2 the results of the architectures such as ResNet18, ResNet50 and others are compared with the results when the proposed SN-BNAM module is integrated with them. Integrating the proposed module with other architectures results in lower error rates with a better params. The proposed SN-BNAM method outperforms other architecture and increases the accuracy of face recognition with reduced error rates. Following this our proposed system is implemented on CIFAR-100 and ImageNet-1K datasets too for effective evaluation. Table 3 and Table 4 lists out the performance evaluation of our SN-BNAM method on CIFAR-100 and ImageNet-1K dataset respectively. The results shows that our method performs better other traditional methods and light weight network architectures in performance and error reduction.

Table 2. Classification results of other network models with our THFDS data set

Architecture	Params	GFLOPs	Top-1 Error (%)	Top-5 Error (%)
ResNet18 [13]	11.22 M	1.81	29.60	10.53
ResNet18+SN-BNAM	11.56 M	1.90	28.36	10.22
ResNet50 [13]	23.11 M	1.22	20.00	7.17
ResNet50+SN-BNAM	23.18 M	1.35	19.97	6.01
ResNet101 [13]	44.23 M	7.65	22.44	6.34
ResNet101+SN-BNAM	45.00 M	7.78	21.41	6.20
ResNeXt29 8x64d [14]	34.43 M	4.99	21.23	9.87
ResNeXt29 8x64d+SN-BNAM	34.52 M	5.12	20.12	9.76
PreResNet110 [15]	1.743 M	0.245	21.89	7.78
PreResNet110+SN-BNAM	1.756 M	0.261	21.54	7.62
VGG-16 [16]	15.231M	7.65	21.96	9.45
VGG-16+SN-BNAM	15.430M	7.71	20.78	9.29
MobileNet [17]	4.33M	0.548	23.47	9.24
MobileNet+SN-BNAM	4.41M	0.566	22.12	9.12

Table 3. Classification results of SN-BNAM on CIFAR-100 dataset

Architecture (CIFAR-100)	Params	GFLOPs	Error
ResNet 50 [13]	23.71M	1.22	21.49
ResNet 50+BAM [9]	24.07M	1.25	20.00
AW-ResNet50 [11]	23.87M	1.23	19.87
ResNet 50+SN-BAM	24.68M	1.28	19.80
ResNet101 [13]	42.07M	2.44	20.00
ResNet101+BAM [9]	43.06M	2.46	19.61
ResNet101+SN-BAM	43.42M	2.29	19.32
PreResNet 110 [15]	1.726M	0.245	22.22
PreResNet 110+BAM [9]	1.733M	0.246	21.96
PreResNet 110+SN-BAM	1.757M	0.249	21.53
WideResNet 28 (w=8) [18]	23.40M	3.36	19.06
WideResNet 28 (w=8)+BAM [9]	23.42M	3.37	19.06
WideResNet 28 (w=8)+SN-BNAM	23.87M	3.41	19.0

Table 4. Classification results of SN-BNAM on ImageNet-1K dataset

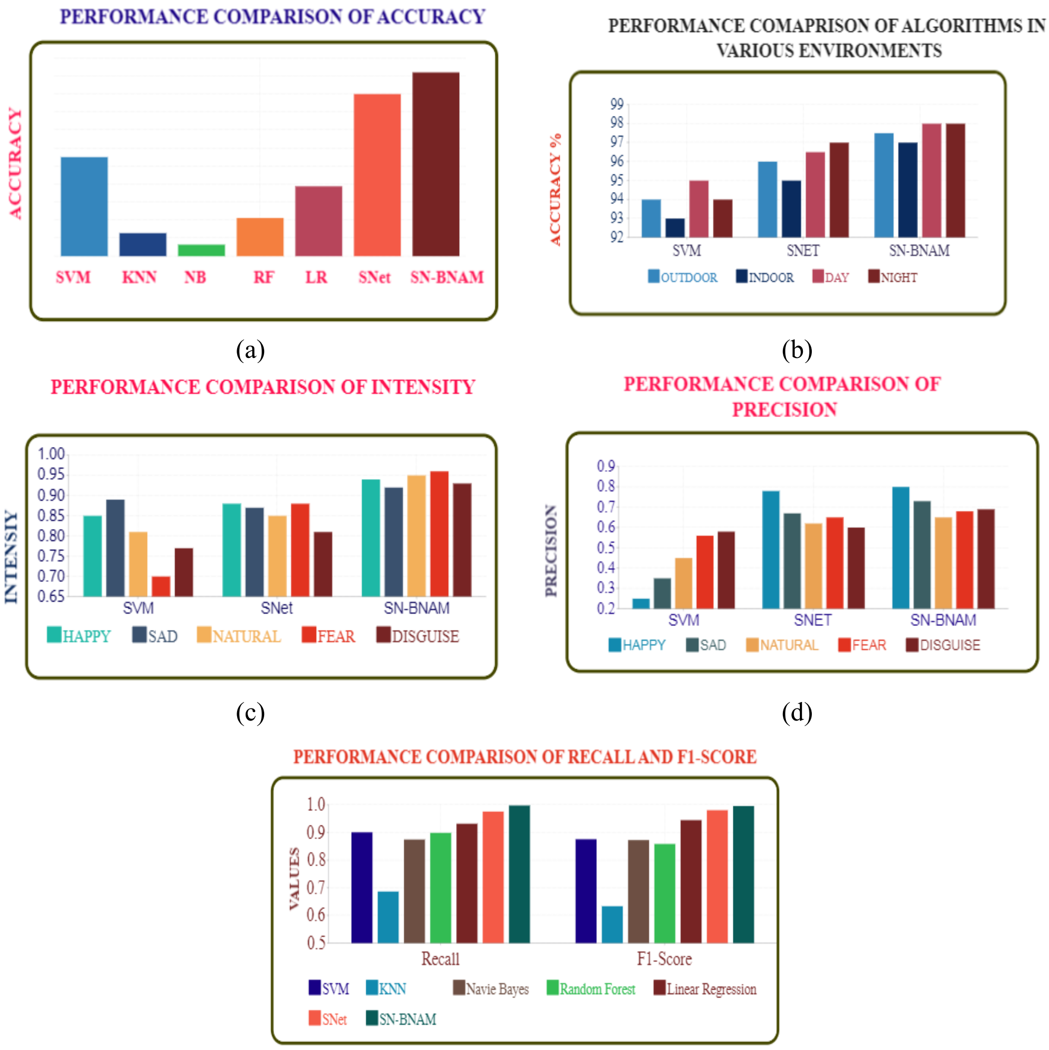
Architecture(ImageNet-1k)	Params	GFLOPs	Top-1 (%)	Top-5 (%)
ResNet 50 [13]	25.56M	3.858	24.56	7.50
ResNet 50+CBAM [12]	28.09M	3.864	22.66	6.31
AW-ResNet50 [11]	25.72M	3.87	23.38	6.79
ResNet 50+BAM [9]	25.92M	3.94	24.02	7.18
ResNet 50+SN-BAM	29.36M	3.897	21.80	5.89
ResNet101 [13]	44.55M	7.570	23.38	6.88
ResNet101+CBAM [12]	49.33M	7.581	21.51	5.69
AW-ResNet101 [11]	44.95M	7.58	22.38	6.21
ResNet101+SN-BAM	49.67M	7.593	20.89	5.34
MobileNet [17]	4.23M	0.569	31.39	11.51
MobileNet+BAM [9]	4.32M	0.59	30.58	10.90
AW-SE-MobileNet [11]	5.52M	0.623	29.41	10.59
MobileNet+SN-BNAM	5.82M	0.687	29.21	10.24
WideResNet 18 [18] (w=1.5)	25.88M	3.866	26.85	8.88
WideResNet 18 (w=1.5)+BAM [9]	25.93M	3.88	26.67	8.69
WideResNet 18 (w=1.5)+CBAM [12]	26.08M	3.868	26.10	8.43
WideResNet18 (w=1.5)+SN-BNAM	26.12M	3.861	25.98	8.12

Table 5. MS COCO object detection [Refer ECCV Paper] (DONE)

Method	Detector	mAP@.5	mAP@.75	mAP@ (0.5, 0.95)
ResNet50 [13]	Faster-RCNN [19]	46.2	28.1	27.0
ResNet50+CBAM [12]	Faster-RCNN [19]	48.2	29.2	28.1
ResNet50+SN-BNAM	Faster-RCNN [19]	49.1	30.1	29.2
ResNet101 [13]	Faster-RCNN [19]	48.4	30.7	29.1
ResNet101+CBAM [12]	Faster-RCNN [19]	50.5	32.6	30.8
ResNet101+SN-BNAM	Faster-RCNN [19]	52.3	33.4	31.7

Table 6. Object detection mAP on the VOC 2007 test set. StairNet detection framework is adopted and SE, CBAM and SN-BNAM are applied to the detectors

BACKBONE	DETECTOR	mAP @.5	PARAMETERS (M)
VGG-16 [16]	SSD [20]	77.8	26.5
VGG-16 [16]	StairNet [21]	78.9	32.0
VGG-16 [16]	StairNet+SE [10]	79.1	32.1
VGG-16 [16]	StairNet+CBAM [10]	79.3	32.1
VGG-16 [16]	StairNet+SN-BNAM	79.5	32.3
MobileNet [20]	SSD [20]	68.1	5.81
MobileNet [20]	StairNet [21]	70.1	5.98
MobileNet [20]	StairNet [21]+SE [10]	70.0	5.99
MobileNet [20]	StairNet [21]+CBAM [12]	70.5	6.00
MobileNet [20]	StairNet [21]+SN-BNAM	70.9	6.04



(a)

(b)

(c)

(d)

(e)

Figure 4. Performance comparison with state-of-art algorithms. (a) Accuracy comparison with other traditional algorithms. (b) Performance of various algorithms in different environmental conditions. (c) & (d) Shows the comparison of algorithms in detecting the 5 basic facial expressions in terms of intensity and precision. (e) The recall and F1-score of various algorithms

4.1 MS COCO object detection

The performance of our proposed system is evaluated by conducting object detection using the Microsoft (MS) COCO dataset [22]. Our model is trained using all the training images, a subset of validation images and 5,000 example images for validation. Faster-RCNN [22] is adopted as our detection method and ImageNet pre trained ResNet101 [15] is chosen as the baseline network. The performance is improved by plugging the SN-BNAM to the baseline architecture. From Table 5, it is observed that our proposed model SN-BNAM shows significant improvements and performs better than other methods.

4.2 VOC 2007 object detection

Experiments were performed on the PASCAL VOC 2007 test set where, SN-BNAM is applied to the detectors. The proposed SN-BNAM method adopts StairNet [23] framework, which is a strongest multiscale method that is based on SSD [24]. SN-BNAM is placed before every classifier to refine the final features prior to the prediction process and enforce the model to select on the features that are meaningful.

The experimental results are listed in Table 6 and it is clearly visible that SN-BNAM has higher accuracy when compared to strong baseline architectures. The improvement in the accuracy is achieved with a slight parameter overhead. The results of the light-weight networks also shows that SN-BNAM can be implemented in low-end devices too.

4.3 Performance comparison with the state-of-art algorithms

The performance of the proposed system is compared with the state-of-art algorithms such as, Support Vector Machine (SVM), K-Nearest Neighbour (KNN), Naïve Bayes (NB), Random Forest (RF), Linear Regression (LR) and SNet architecture. The results are displayed in Figure 4. The performance comparisons are done using the metrics such as accuracy, intensity, precision, recall and F1-score. Experiments were done on various environments like indoor, outdoor, day and night and it is clear that the proposed SN-BNAM method outperforms other state-of-art algorithms. The intensity and precision are reported for the five basic face expressions namely, happy, sad, natural, fear and disguise. Our proposed method performs better in all aspects when compared to all other algorithms.

5. CONCLUSION

In this paper, a SENet architecture embedded with BottleNeck Attention module is proposed for face recognition. The proposed SN-BNAM method uses single thermal face images of human to overcome the limitations of visible light images or RGB images, such as illusion, occlusion, pose variations and variable lighting conditions. The problem of aging, images of person wearing spectacles or mask has no effect on the performance accuracy while using thermal face images. Moreover, along with face recognition, the proposed system also detects the six basic face expressions of human in any time of day such as indoor, outdoor, day and night. Our method works by capturing the important features in the input image. The proposed method chooses SNet architecture as its backbone architecture to get better accuracy when compared to other baseline architectures. The input thermal face image is processed by the SE block and BAM block which is kept between the BottleNeck locations. The Channel and the spatial attentions inside the BAM perform the work of what and where to focus. These attention modules focus only on the important features for face recognition and other unimportant features are left behind. The experiments done in this paper clearly demonstrate that SN-BNAM improves the baseline architectures like MobileNet, VGG-16, ResNet and others on image classification of our own THFDS data set and also on the CFIAR-100 and ImageNet dataset. The proposed system works well object detection on MS COCO and VOC 2007 data set and the results are reported. A comparison with the traditional predicting algorithms such as SVM, KNN, Decision Trees are also done and our SN-BNAM method out performs all in accuracy and error reduction. By integrating the proposed SN-BNAM module with other baseline architectures the performance of other architecture gets elevated in terms of accuracy and reduced error rates. The channel and spatial attention improve the accuracy by removing noise in the input image and focus is given only on the feature attributes. The proposed system uses only single human face image for experiment. As a future work, the dataset can be improvised by including full size images of persons and videos and methods and algorithms can be developed to select or crop only the face of the human from the video for facial recognition.

REFERENCES

[1] Vigneau, G.H., Verdugo, J.L., Castro, G.F., Pizarro, F., Vera, E. (2017). Thermal face recognition under temporal variation conditions. Ieee Access, 5: 9663-9672. https://doi.org/10.1109/ACCESS.2017.2704296

[2] Weidlich, V.A. (2021). Thermal infrared face recognition. Cureus, 13(3). https://doi.org/10.7759/cureus.13736

[3] Dixit, A.N., Kasbe, T. (2020). A survey on facial expression recognition using machine learning techniques. 2nd International Conference on Data, Engineering and Applications, 2020, pp. 1-6. https://doi.org 10.1109/IDEA49133.2020.9170706

[4] Vargas, R., Mosavi, A., Ruiz, R. (2018). Deep learning: A review. Preprints.org, 2018: 2018100218. https://doi.org/10.20944/preprints201810.0218.v1

[5] Pei, Z., Xu, H., Zhang, Y.N., Guo, M., Yang, Y.H. (2019). Face recognition via deep learning using data augmentation based on orthogonal experiments. Electronics, 8(10): 1088. https://doi.org/10.3390/electronics8101088

[6] Hammadi, O.I., Abas, A.D., Ayed, K.H. (2018). Face recognition using deep learning methods a review. International Journal of Engineering and Technology, 7(4): 6181-6188.

[7] Misra, D., Nalamada, T., Arasanipalai, A.U., Hou, Q. (2021). Rotate to attend: Convolutional triplet attention module. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 3139-3148. https://doi.org/10.1109/WACV48630.2021.00318

[8] Zhang, Z.X., Wang, M.W. (2022). Convolutional neural network with convolutional block attention module for

finger vein recognition. arXiv preprint arXiv: 2202.06673. https://doi.org/10.48550/arXiv.2202.06673

[9] Park, J., Woo, S., Lee, J.Y., Kweon, I.S. (2018). Bam: BottleNeck attention module. arXiv preprint arXiv: 1807.06514. https://doi.org/10.48550/arXiv.1807.06514

[10] 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. https://doi.org/10.1109/CVPR.2018.00745

[11] Zhu, B.Z., Hofstee, P., Lee, J., Al-Ars, Z. (2021). An attention module for convolutional neural networks. arXiv preprint arXiv: 2108.08205. https://doi.org/10.48550/arXiv.2108.08205

[12] Woo, S., Park, J., Lee, J.Y., Kweon, I.S. (2018). Cbam: Convolutional block attention module. In Proceedings of the European conference on computer vision (ECCV), pp. 3-19. https://doi.org/10.1007/978-3-030-01234-2_1

[13] He, K., Zhang, X.Y., Ren, S.Q., Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770-778. https://doi.org/10.1109/CVPR.2016.90

[14] Xie, S., Girshick, R., Dollár, P., Tu, Z.W., He, K. (2017). Aggregated residual transformations for deep neural networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1492-1500. https://doi.org/10.1109/CVPR.2017.634

[15] He, K.M., Zhang, X.Y., Ren, S.Q., Sun, J. (2016). Identity mappings in deep residual networks. In Computer Vision-ECCV 2016: 14th European Conference, Springer International Publishing, pp. 630-645. https://doi.org/10.1007/978-3-319-46493-0_38

[16] Simonyan, K., Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv: 1409.1556. https://doi.org/10.48550/arXiv.1409.1556

[17] Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T. Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv: 1704.04861. https://doi.org/10.48550/arXiv.1704.04861

[18] Zagoruyko, S., Komodakis, N. (2016). Wide residual networks. arXiv preprint arXiv: 1605.07146. https://doi.org/10.48550/arXiv.1605.07146

[19] Ren, S., He, K., Girshick, R., Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. In IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(6): 1137-1149. https://doi.org/10.1109/TPAMI.2016.2577031

[20] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C. (2016). Ssd: Single shot multibox detector. In Computer Vision-ECCV 2016: 14th European Conference, Springer International Publishing, pp. 21-37. https://doi.org/10.1007/978-3-319-46448-0_2

[21] Woo, S., Hwang, S., Kweon, I.S. (2018). Stairnet: Top-down semantic aggregation for accurate one shot detection. In 2018 IEEE winter conference on applications of computer vision (WACV), IEEE, pp. 1093-1102. https://doi.org/10.1109/WACV.2018.00125

[22] Ren, S., He, K., Girshick, R., Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. In Proc. of Neural Information Processing Systems (NIPS), 2015.

[23] Park, J., Woo, S., Lee, J.Y. (2018). BAM: Bottleneck attention module. In 2018 Computer Vision and Pattern Recognition. https://doi.org/10.48550/arXiv.1807.06514

[24] Zhou, S., Qiu, J. (2021). Enhanced SSD with interactive multi-scale attention features for object detection. Multimedia Tools and Applications, 80: 11539-11556. https://doi.org/10.1007/s11042-020-10191-2