Neural Computing and Applications (2023) 35:13321–13341
https://doi.org/10.1007/s00521-023-08440-0

ORIGINAL ARTICLE

Micro-network-based deep convolutional neural network for human activity recognition from realistic and multi-view visual data

Arati Kushwaha[1] · Ashish Khare[1] · Om Prakash[2]

Received: 15 January 2022 / Accepted: 24 February 2023 / Published online: 13 March 2023

Abstract

In the recent past, deep convolutional neural network (DCNN) has been used in majority of state-of-the-art methods due to its remarkable performance in number of computer vision applications. However, DCNN are computationally expensive and requires more resources as well as computational time. Also, deeper architectures are prone to overfitting problem, while small-size dataset is used. To address these limitations, we propose a simple and computationally efficient deep convolutional neural network (DCNN) architecture based on the concept multiscale processing for human activity recognition. We increased the width and depth of the network by carefully crafting the design of network, which results in improved utilization of computational resources. First, we designed a small micro-network with varying receptive field size convolutional kernels (1×1, 3×3, and 5×5) for extraction of unique discriminative information of human objects having variations in object size, pose, orientation, and view. Then, the proposed DCNN architecture is designed by stacking repeated building blocks of small micro-networks with same topology. Here, we factorize the larger convolutional operation in stack of smaller convolutional operations to make the network computationally efficient. The softmax classifier is used for activity classification. Advantage of the proposed architecture over standard deep architectures is its computational efficiency and flexibility to use with both small as well as large size datasets. To evaluate the effectiveness of the proposed architecture, several extensive experiments are conducted by using publically available datasets, namely UCF sports, IXMAS, YouTube, TV-HI, HMDB51, and UCF101 datasets. The activity recognition results have shown outperformance of the proposed method over other existing state-of-the-art methods.

Keywords Convolutional neural network · Human activity recognition · Micro-network · Softmax classifier

1 Introduction

With the rapid development of digital media technology such as surveillance, film crew and mobile phone, computer vision scientists have increased interest in development of automated monitoring system. Therefore, vision-based human activity recognition (HAR) has become one of the most prominent research area due to its numerous applications in intelligent security monitoring, entertainment, smart indoor security, military applications, healthcare, robot vision, day-to-day activity monitoring [1, 2], etc. HAR system aims to automate the video monitoring system to help the human operator in identifying unusual events of interest. A number of works have already been done in this area with significant improvement in accuracy but accurate activity recognition is still a challenging task [2]. In the past decade, a large number of researches have given the methods for human activity recognition that use different handcrafted features [2–7] such as histogram of oriented gradient (HOG) [3], local binary pattern (LBP) [4], local ternary pattern (LTP) [5], scale-invariant feature transform (SIFT) [6], Harris3D [7], etc. The methods based on handcrafted features achieved success up to certain

✉ Ashish Khare
khare@allduniv.ac.in

Arati Kushwaha
aratikushwaha.jk@gmail.com

Om Prakash
au.omprakash@gmail.com

[1] Department of Electronics & Communication, University of Allahabad, Prayagraj, Uttar Pradesh, India

[2] Department of Computer Science & Engineering, HNB Garhwal University, Srinagar Garhwal, India

extent for the videos captured in controlled environments. But, the challenges of accurate human activity recognition still lies for real-world applications since realistic videos are complex in nature and have a dynamic range of varying information. Also in real time applications, it is difficult to decide which feature will be suitable for the problem at hand. A small variation in motion, scale and object pose can generate similar feature values in different categories of activity classes and different feature values in same category of activities which may lead to poor classification [8].

In recent past, deep learning-based approaches have outpaced the handcrafted feature-based conventional approaches due to its success in number of computer vision applications [9–16]. The self-learning capability of deep learning networks from complex representation of visual data may help deep learning architectures suitable for video-based human activity recognition [10]. After the success of AlexNet, several deep architectures have been considered for computer vision applications with the aim to achieve better performance in a limited computational cost. The most straightforward way of improving classification accuracy is to increase size of the network in terms of network depth and width. It has been studied by the researchers that deeper architectures can grab dynamic range of complex details from complex visual data than the shallower ones [13]. However, deeper architectures need large number of learnable parameters and plenty of computational resources for training. These architectures suffer from overfitting problem with smaller size datasets.

Enormous works have been done on human activity recognition based on deep learning. Researchers working in this area have used fusion of two networks, integration of handcrafted features and deep learning architectures, and 3D CNN-based architectures to achieve better performance which came true up to certain extent. But with the advancements in mobile computing devices and robotics, design of an efficient algorithm is still needed that performs better in limited computational budget.

Therefore, we proposed a simple and computationally efficient deep convolutional neural network (CNN) architecture for human activity recognition. The proposed architecture is constructed by stacking the repeated building blocks (small micro-networks) of same topology. The micro-networks are small CNN architectures designed to cluster the neurons, and their outputs are highly correlated at each layer. Micro-networks are constructed using convolutional kernels with varying receptive fields. The designed architecture will grab dynamic range of complex details for each activity category from the complex visual data that have large variations in scale and poses of human objects.

The main contributions of the proposed work are as follows:

(i) We designed a simple and computationally efficient deep CNN architecture based on small micro-networks that have lesser number of hyperparameters than the standard deep learning architectures, and it can also be trained on low computing devices or scenarios that have inherently limited computational budget such as mobile vision technologies.

(ii) The proposed network is fine-tuned and trained from scratch using raw RGB data then evaluated using a softmax classifier.

(iii) Several extensive experiments have been performed to validate the authenticity of the proposed network. To establish the soundness of the proposed architecture, compared it with its close variants in terms of learnable parameters and convergence rate.

To validate the performance of the proposed framework of human activity recognition, we conducted experiments on six different publically available human activity recognition video datasets and compared the results with several existing state-of-the-art methods. The experimental results have demonstrated the usefulness and effectiveness of the proposed human activity recognition method.

Rest of the paper is organized as follows: The details of related work is given in Sect. 2. The description of the proposed work for human activity recognition is given in Sect. 3. In Sect. 4, we presented the experimental setup and datasets considered in the proposed work. The experimental results are discussed in Sect. 5, and finally Sect. 6 concludes the paper.

2 Related work

Video-based human activity recognition is a difficult task due to several challenges like fuzzy boundary between activity categories, varying view-point, inter- and intra-class variations, similarity between different categories of activity, object occlusion, varying illumination conditions, camera motion, presence of noise, cluttered background, non-rigid human object and ambiguous definition of different actions [1, 2], etc. Selection and extraction of suitable features play a vital role in activity recognition task. Good discriminative features enhance the performance, while poor and ambiguous features degrade the performance of activity recognition. Based on feature extraction techniques, the literature related to human activity recognition is categorized into two categories, namely conventional handcrafted feature-based approaches and deep learning-based approaches.

In the past decade, a number of handcrafted feature descriptors have been exploited by researchers such as [3–7], etc., for human activity recognition. Based on the combination of optical flow vectors and histogram of oriented magnitude a novel feature descriptor have been proposed by Arati et al. [2] for activity recognition. Alina et al. [17] have developed a framework for human activity recognition using skeleton data in which they used a random forest classifier for activity recognition. Arati et al. [18] proposed a framework for human activity recognition in which they used multiple features in order to uniquely represent complex information for each activity category. They constructed the feature vector based on integration of Discrete Wavelet Transform, Multiclass LBP, and HOG features and then used one-vs-one multiclass support vector machine for activity recognition. Roshan at al. [19] have presented human activity recognition framework based on combination of multiple handcrafted feature representation techniques for multi-view environment and then used hidden Markov model for activity recognition. In [20], Swati et al. proposed a framework for human activity recognition for video sequences in which they used an integration of moment invariants and uniform local binary patterns followed by multiclass SVM. Muhammad et al. [21] have considered a hybrid approach based on multiple features to extract feature vectors and then used rank correlation-based feature selection approach for selecting appropriate features followed by KNN multiclass classifier for activity recognition. Hand-crafted feature based approaches have achieved success up to certain extent but still there is a need to design algorithms for realistic videos recorded in complex uncontrolled environments.

In recent years, deep learning-based models have become a mainstream method for computer vision applications [8–14, 17, 22, 23]. Motivated by this, several researchers have published their work on human activity recognition based on deep learning architectures [8, 24–29]. In [8], a resource-conscious deep learning architecture which consists of total 26 layers has been proposed by Muhammad et al. for vision-based human activity recognition. They used a statistical approach for unique unambiguous feature selection based on Poisson distribution followed by softmax classifier for action recognition. A 3D asymmetric MicroNets based method for human action recognition has been proposed by Hao et al. [24] in which they used several MicroNets to incorporate the multiscale processing. Noor et al. [25] have proposed a framework for human action recognition in which they used a video summarization technique followed by 3D deep CNN architecture. Muhammad et al. [26] have proposed a framework for human action recognition, in which they computed deep learning features by pre-trained VGG-16 model and handcrafted feature by using horizontal and vertical gradients followed by feature fusion strategy to construct the feature vector. The final feature vector for activity recognition was constructed by selecting high probable features based on the three parameters—relative entropy, mutual information and strong correlation coefficient (SCC). Tran et al. [27] have proposed a 3D deep CNN architecture to grab spatiotemporal features to achieve significant improvement in accuracy value for human action recognition. Sachin et al. [28] have proposed a deep CNN architecture for human action recognition in which they first computed depth images and then these depth images are used for training and testing purposes. In [29], Mei et al. have proposed a semi-CNN architecture based on the concept of fusion of 2D and 3D CNN architectures to encode spatiotemporal information for human action recognition.

From the above detailed literature review, we found that several approaches, based on conventional features as well as deep learning, for video-based activity recognition exists. Although a number of work have been done for human activity recognition and have achieved remarkable success in terms of classification accuracy, people are still trying to develop efficient algorithms which can work well in limited computational budget with increased performance. Therefore in this work, we proposed a computationally efficient deep CNN architecture based on micro-networks for human activity recognition that have lesser number of parameters and could be trained on low computing devices.

3 The proposed method

The ultimate goal of the proposed work is to introduce a simple and computationally efficient CNN architecture which works well in limited computational budget and has flexibility of training with small and large size datasets, with improved performance. In this work, we propose a supervised learning-based multiscale architecture for human activity recognition that has the capability to learn complex invariant features from realistic video data and deals with challenges of varying size of objects, varying object poses and various image transforms. The proposed approach consists of the following main steps:

(i) Collect large video data and resize them using augmentation techniques before feeding for network training and to avoid the overfitting problem also.

(ii) Design small micro-networks that have varying convolutional kernels on the same layer to process data using a combination of convolutional, ReLU and batch normalization layers. This design provides multiscale processing.

(iii) Design a simple and optimized CNN architecture by stacking repeating building blocks (stacking small micro-networks) with the same network topology.

(iv) Fine-tune the proposed network and train the proposed network from scratch using raw RGB data and evaluate the trained network using softmax classifier after training.

3.1 General design principle of the proposed architecture

Although several works have been done for activity recognition based on deep learning methods, selection of an optimum deep learning architecture is still a difficult task and is application-dependent [25]. Further, it has been proven that the deeper architectures have better generalization ability and are able to learn more discriminative features hierarchically. The most straightforward way of increasing the size of the network is by increasing the depth and width of the network on each layer. But by simply stacking convolutional layers to design the deep architecture makes the algorithm computationally expensive. Therefore, such deeper architectures are not suitable for mobile vision devices that have limited computing capability and constrained memory. Also by uniformly increasing the network size, it become prone to overfitting problem with smaller size dataset. Thus, the need is to carefully design the CNN architecture with an increase in depth and width of the network.

Further, it has been studied that visual data of human activities recorded in realistic environment consists of dynamic range of complex information due to complex human motions, which lead to the challenges like large inter-class variations in the same activity category and fuzzy boundaries between different activity categories caused due to variations in scale, pose, and viewpoint changes as illustrated in Fig. 1.

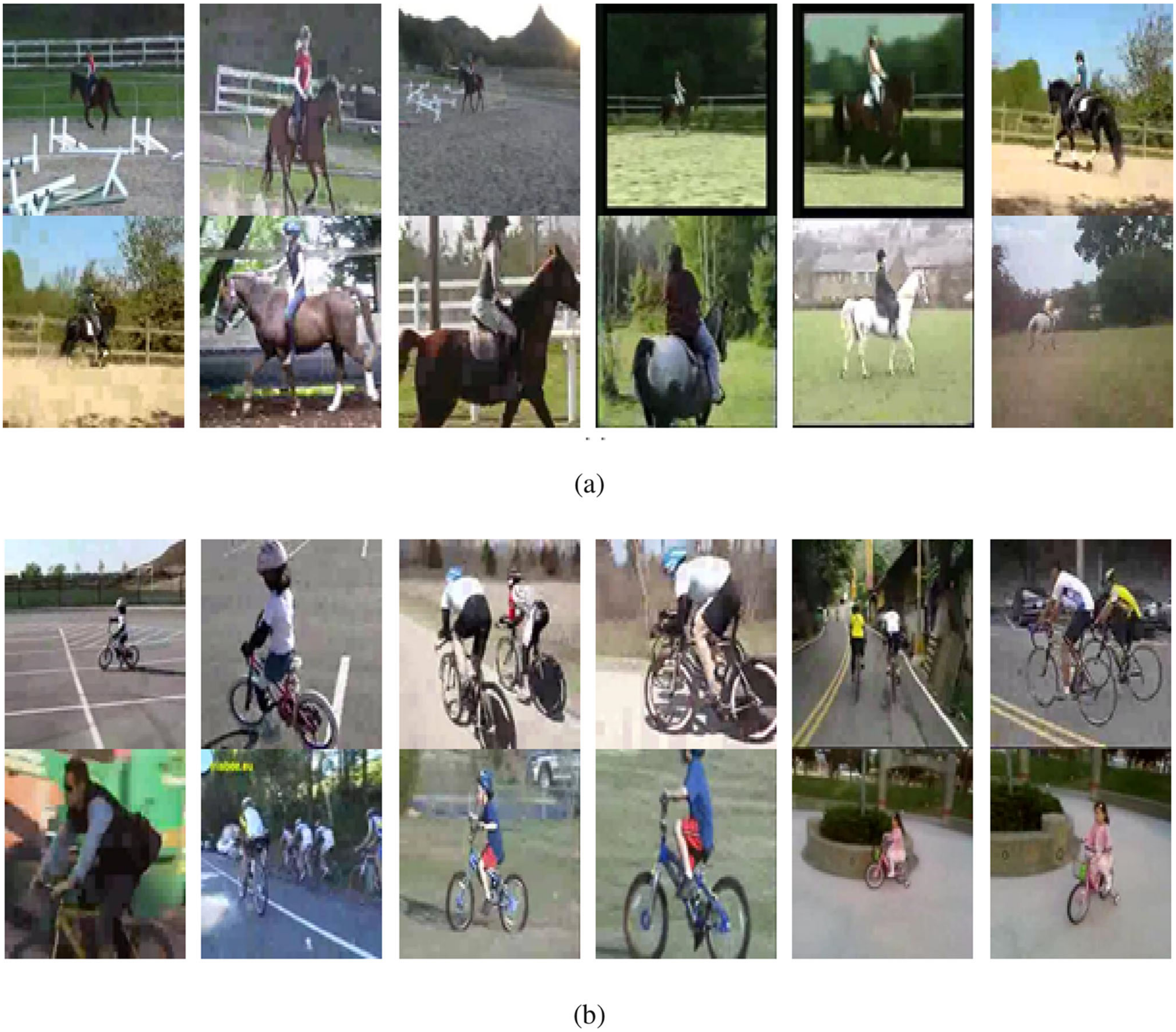
Figure 1 shows that varying object size, pose, orientation and views of human objects in the sample frames represent different activity categories. Therefore, unique discrimination of each activity category requires several local and global structural information of each activity category. Thus, during the design of CNN architecture for human activity recognition, choice of right size convolutional kernel is difficult due to large variation in distribution of information across sample frames of each activity category. We can overcome the above-mentioned challenge up to certain extent by designing deep architecture with varying receptive field size convolutional kernels at a particular layer, which can encapsulate the dynamic range of complex patterns of human activities that have variations in scale, orientation, and pose. A larger size convolutional kernel can be used to capture information which is more spread out in the frame and a smaller size convolutional kernel can be used for information which is less spread out [11].

Inspired by the method proposed by Christian et al. [11], we proposed a deep CNN architecture based on the concept of multiscale processing in which we used varying size convolutional kernels at a same layer of network. Motivated by the works presented in [11, 24], we designed a small micro-network with varying size convolutional kernels (1×1, 3×3, and 5×5) as shown in Fig. 2a. The proposed deep convolutional neural network architecture is constructed by stacking repeated building blocks of these small micro-networks. The micro-network is used in this work to increase the depth and width of the network simultaneously and to enhance the learning capability of the network without increasing the computational budget. The proposed architecture is deeper and wider than the standard deep learning architectures and has the capability to get trained on low-memory GPU devices. The proposed architecture has the potential to process complex patterns at multiple scale which helps in robust discrimination of each activity category uniquely and makes the network learning process faster.

3.2 Factorizing convolutional operation with smaller filters

The computational efficiency and lesser number of learning parameters are essential factors in designing of deep CNN architecture for low computing devices. Therefore, for efficient utilization of computational resources of the system, we further factorize the larger convolutional operation of the micro-network into a smaller size convolutional operation in a manner that have the same effect on the receptive field size of the larger convolution operation of the network as illustrated in Fig. 3. Figure 3 represents decomposition of 5×5 convolutional operation using stack of two 3×3 convolutional operations.

Thus, to increase the computational speed, the convolutional layer C3 in path 3 (Fig. 2a) with convolutional kernel of size 5×5 is replaced by two 3×3 convolutional kernels (as shown in Fig. 2b). This leads to the reduction in number of learnable parameters, i.e., stacks of two convolutional layers with 3×3 kernels along with C channels requires $2 \times (3^2C^2) = 18C^2$ parameters, whereas single convolutional layer with 5×5 kernel size needs $5^2C^22 = 25C^2$ parameters. Stacking of two convolutional layers



(a)

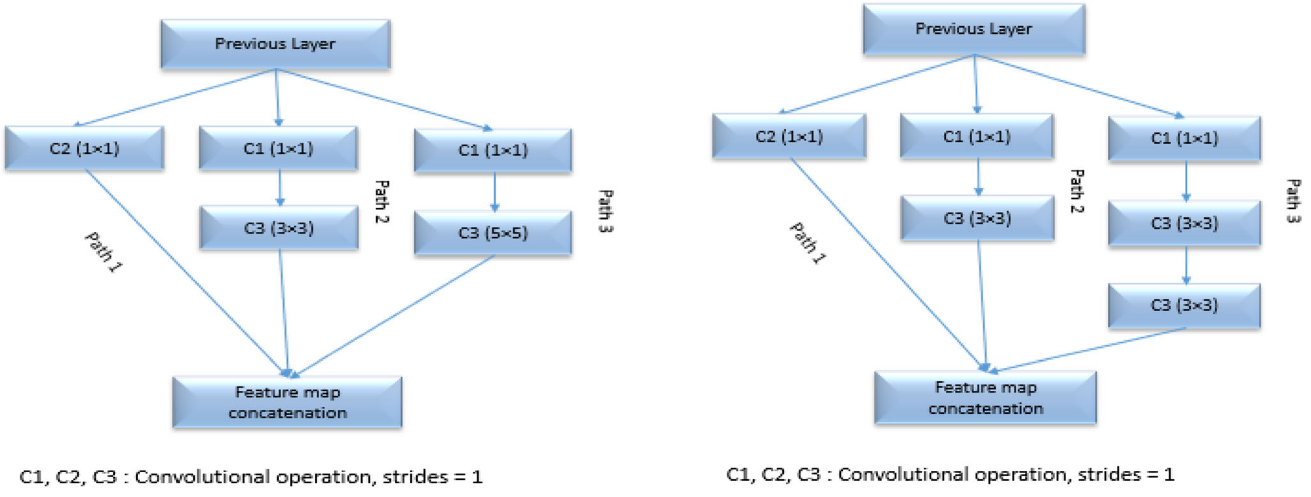
(b)

Fig. 1 Sample frames of UCF101 dataset [19] from two activity categories **a** Horse riding and **b** Cycling

with kernel size 3×3 instead of a single convolutional layer with 5×5 increases the depth of network which introduces more nonlinearity in the network [30].The merit of stacking two convolutional layer (of size 3×3) instead of single 5×5 convolutional layer is that smaller size filter helps in extracting fine-grained details of activity data. And also increasing depth of the network allows network to learn more complex details. Therefore, the micro-network presented in Fig. 2b is used to design the CNN for the proposed work.

3.3 Architectural detail

Inspired from inception modules proposed in [11, 24], we designed small micro-networks having multiple size convolutional kernel in which larger size kernel is for capturing the globally distributed information and smaller size kernel is for capturing the locally distributed information. Results obtained after applying all convolutional kernels on a particular level are concatenated and used as an input to the next level. The proposed micro-network is shown in Fig. 2b. The proposed micro-network is a CNN architecture that is constructed with C1, C2, and C3 convolutional layers with 1×1, 3×3 and 5×5 (equivalently stack of two convolution operations of kernel size 3×3) convolutional kernels followed by ReLU and batch normalization layer to process input feature at multiple scales [9]. The convolutional layer C1 in path 2 and path 3 used kernel size 1×1 before 3×3 and 5×5 operation to reduce the dimensionality of the channel before passing through



(a) Micro-network-A

(b) Micro-network-B

Fig. 2 Proposed micro-networks

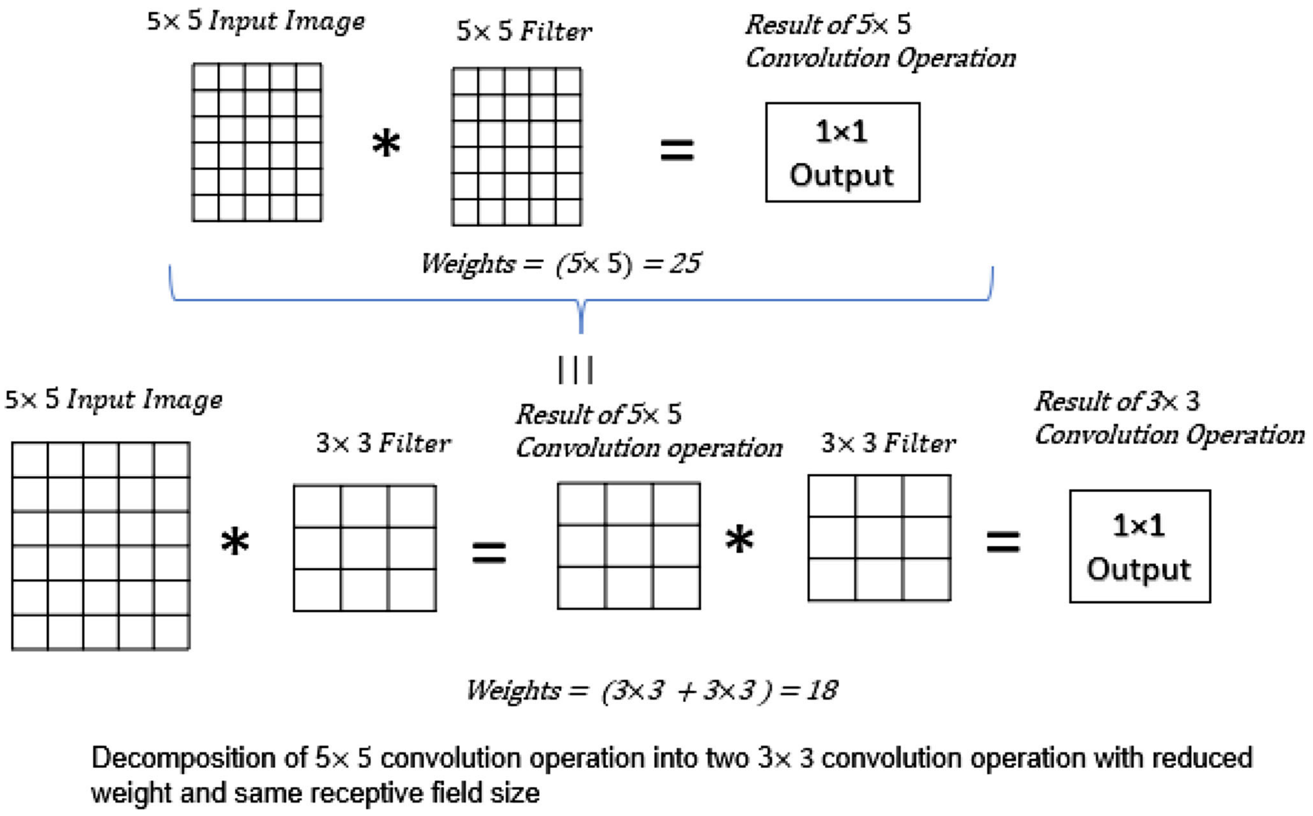

5× 5 Input Image
5× 5 Filter
Result of 5× 5 Convolution Operation
1×1 Output
Weights = (5× 5) = 25
|||
5× 5 Input Image
3× 3 Filter
Result of 5× 5 Convolution operation
3× 3 Filter
Result of 3× 3 Convolution Operation
1×1 Output
Weights = (3×3 + 3×3) = 18
Decomposition of 5× 5 convolution operation into two 3× 3 convolution operation with reduced weight and same receptive field size

Fig. 3 Illustration of replacing 5×5 convolutional operation using stack of two 3×3 convolutional operations

the network. This reduces the computational complexity of the network and also increases the width and depth of the network. The final output feature map of path 1, path 2, and path 3 are concatenated and used as an input to the next layer. The output at layer of this micro-network is mathematically represented as follows:

$$f_l = T_l([f_1, f_2, f_3]) \quad (1)$$

where $[f_1, f_2, f_3]$ refers to concatenation of feature maps and $T_l(.)$ is nonlinear transformation.

Thus, with the carefully crafted design of small micro-network, we increased the network depth and width and that too in a limited computational budget with multiscale processing capability. Varying size convolutional kernel on a layer of the proposed architecture is used to extract various feature maps to capture different complex patterns of activity data. These feature maps extracted with different varying convolutional operations are concatenated and used as input to the next layer. The proposed deep learning architecture for human activity recognition is shown in Fig. 4.

This architecture is constructed by stacking micro-networks after three convolutional layers followed by one fully connected layer with 256 units and a softmax classifier. Each convolutional layer of the proposed architecture and micro-network is followed by ReLU activation function and batch normalization to introduce nonlinearity in the network and for generalization of the network, to enhance the discriminative power of the decision function and speed up the learning process [9, 31]. In the proposed architecture, first two micro-networks consist of 64, (64, 96), (8, 16, 16) convolutional kernels, the next two micro-networks have 96, (96,128), (16, 32, 32) convolutional kernels, and the last two micro-networks have 128, (128,256), (32, 64, 64) convolutional kernels. The proposed architecture also contains four max-pooling layers (M1-M4) with window size (3,3) and stride (2,2), one average pooling layer A1 with window size (5,5), and stride (1,1). Table 1 presents the detailed architectural description of the proposed architecture.

Human activity data have a dynamic range of complex information. Similar activities can have extremely large variations in the size of human objects. Therefore, choosing the right size of convolutional kernel is important. With the larger convolutional kernel size, we can grab information that is distributed in large area and smaller size of convolutional kernel help to grab locally distributed information. Therefore, we used different size convolutional kernels in a micro-network layer which helps in multiscale processing of dynamic range of information. Therefore, the proposed deep CNN architecture with multiscale processing in micro-networks has capability to deal with the challenges of large variations in sizes, orientations, and poses of human objects within the same activity class.

4 Experiments and datasets used

This section presents an elaboration of implementation details, evaluation criteria and datasets used in our experiments.

4.1 Implementation detail and evaluation criteria

We trained the proposed architecture from scratch on several publically available datasets discussed in Sect. 4.2. To implement the proposed architecture, we used Python3, Keras and TensorFlow deep learning libraries. All the experiments have been performed on NVIDIA P2000 GPU having Intel® Xeon® CPU E7- 4809 processor. The proposed architecture was trained from scratch using stochastic gradient descent (SGD) optimizer with initial learning rate 0.01, batch size 64, momentum 0.9, and weight decay 1e-4 [32]. We have also considered close

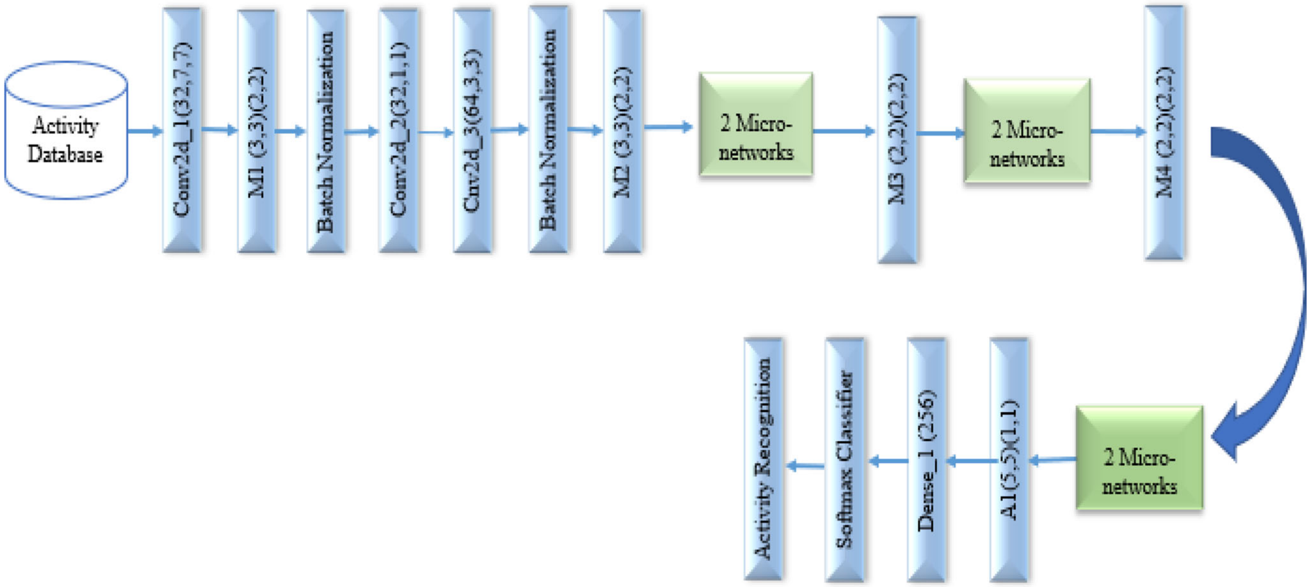


Fig. 4 Proposed deep learning architecture for human activity recognition

variants of the proposed architecture by using variants of the proposed micro-network (bottleneck_1 and bottleneck_2) as illustrated in Fig. 5a and b.

The bottleneck shown in Fig. 5a is proposed in [11]. Learnable parameters, convergence rate, and classification accuracy are used as performance measure to evaluate the effectiveness of the proposed architecture [13]. The proposed architecture has also been evaluated with different initial learning rate.

4.2 Dataset description

The soundness of the proposed architecture is tested by conducting several extensive experiments on six publically available benchmark human activity datasets UCF sports [33], IXMAS [34], YouTube [35], TV-HI (TV Human Interaction) dataset [36], HMDB51 [37] and UCF101 [38]. The sample frames of the considered datasets are shown in Fig. 6.

Table 1 Architectural details of the proposed network

Layer	Kernel size	Proposed architecture	Output size
Convolution	7×7	32, Conv, stride 2, ReLU, BN	75×75
Pooling	3×3	Max pooling, stride 2	38×38
Convolution	1×1	32, Conv, stride 1, ReLU	38×38
Convolution	3×3	64, Conv, stride 1, ReLU, BN	38×38
Pooling	3×3	Max pooling, stride 2	19×19
Micro-network	–	$\begin{bmatrix} 1\times 1Conv \\ 1\times 1, 3\times 3Conv \\ 1\times 1, 3\times 3, 3\times 3Conv \end{bmatrix}\times 2$	19×19
Pooling	3×3	Max pooling, stride 2	10×10
Micro-network	–	$\begin{bmatrix} 1\times 1Conv \\ 1\times 1, 3\times 3Conv \\ 1\times 1, 3\times 3, 3\times 3Conv \end{bmatrix}\times 2$	10×10
Pooling	3×3	Max pooling, stride 2	5×5
Micro-network	–	$\begin{bmatrix} 1\times 1Conv \\ 1\times 1, 3\times 3Conv \\ 1\times 1, 3\times 3, 3\times 3Conv \end{bmatrix}\times 2$	5×5
Pooling	5×5	Average pooling, stride 2	1×1
FC	–	256, ReLU	–
Classification layer	–	Softmax Classifier	–

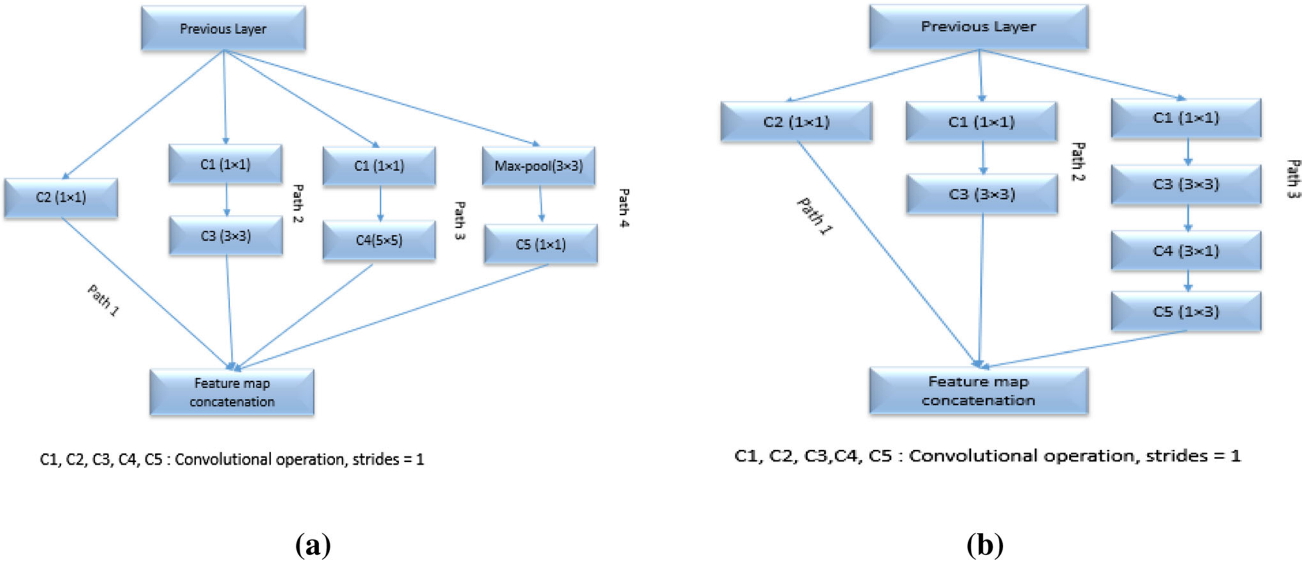


Fig. 5 Close variants of the proposed micro-network **a** Bottleneck_1 [11] **b** Bottleneck_2

4.2.1 UCF sports dataset

UCF sports [33] is a publically available dataset that consists of several sports activity. The videos in this dataset were collected from broadcasted television channels such as BBC and ESPN that are captured in a realistic environment. It consists of total ten activity categories (golf swing, kicking, lifting, diving, horse-riding, running, skateboarding, Swing_Bench, Swing_SideAngle and walking). This dataset consists of a total of 182 videos of resolution 720×480, recorded in real sports environment with challenges like varying lighting condition, complex background, occlusion, etc.

4.2.2 IXMAS dataset

IXMAS [34] is a publically available multi-view activity dataset with total 13 activity classes (check watch, crossing arms, do nothing, getting up, scratching head, sitting down, walking, waving, pointing, punching, picking up, throwing from bottom-up and throwing overhead) performed by 11 peoples. It is a collection of total 1148 low-resolution videos captured by five different cameras from different views under a controlled environment with a 23 fps frame rate and 320×291 revolution.

4.2.3 YouTube dataset

YouTube [35] is a realistic human activity video dataset collected from YouTube. It consists of total 11 activity categories (basketball playing, biking/cycling, diving, golf swing, juggling, horse riding, trampoline jumping, volleyball spiking, walking with dog, and swinging. It is one of the popularly used sports dataset having challenges like object appearance, illumination conditions, viewpoint, object scale and camera motion. Each category of activities of this dataset has been divided into 25 groups with some shared properties in each group like similar background, view point, same human object, etc.

4.2.4 TV-HI dataset

TV-HI (TV Human Interaction) dataset [36] was created in 2012. The main objective of this dataset was to address the recognition of interaction between two human objects. It consists of four types of human interaction activities (handshaking, high-fives, hugs and kisses) with the challenges like clutter background, camera motion, changes in viewpoint, complex background and varying number of people in the scene. This dataset was collected from 20 different TV shows, having a total of 300 video clips.

4.2.5 HMDB51 dataset

HMDB51 [37] is one of the most popular challenging dataset in which videos were collected from YouTube and movies. It consists of total 7K videos with 51 activity categories in which each activity category have minimum 101 video clips with 320×240 resolution and 30 fps frame rate. This dataset consists of activity categories ranging from daily life activities to sports activities, with complex high-level activities and having challenges like complex background, camera motion, varying lighting condition, human–human interaction, and human–object interaction, etc. These 51 activity categories could be categorized into five main groups that are general facial actions (laugh, chew, smile, talk, etc.), facial actions with object manipulation (eat, drink, smoke), general body movements (cartwheel, jump, pull up, push up, run, sit down, stand up, turn, walk, wave, etc.), body movements with object interaction (brush hair, catch, draw sword, dribble, golf, hit something, kickball, pick, throw, etc.) and body movements for human interaction (fencing, hug, kick someone, kiss, sword fight, etc.).

4.2.6 UCF101 dataset

UCF101 [38] is a realistic user uploaded human activity video dataset that have a total of 13320 instances that are recorded under an uncontrolled environment and have challenges like varying illumination conditions, contain partial occlusions, different pose and orientations of human objects, cluttered scenes, complex background, camera motion, etc. The average length of video clips in this dataset is 170 frames per video and duration is about 7 second with 320×240 resolution and 30 fps frame rate. It is one of the most challenging human activity recognition dataset due to the large number of activity categories, the large number of video clips, and also unconstrained nature of such video clips. Its 101 activity categories can be divided into five groups that are human–object interaction, body-motion only, human–human interaction, playing musical instruments and sports.

5 Results and discussion

Using the proposed architecture, we experimented for HAR on several publically available datasets. The datasets used in our experiments have their own challenging situations and actions. The recognition results were recorded for different learning rates, epochs, etc. Learnable parameters, classification accuracy and convergence rate are taken as performance measures to evaluate the results of the proposed architecture. Three sets of experiments were

(a)

(b)

(c)

(d)

(e)

(f)

◀**Fig. 6** Sample frames of the datasets taken for experimentation. **a** UCF sports [33] **b** IXMAS [34] **c** YouTube [35], **d** TV-HI [36], **e** HMDB51 [37], and **f** UCF101 [38]

conducted in order to evaluate the effectiveness and efficiency of the proposed architecture. Firstly, we conducted experiments by training proposed architecture from scratch using RGB data using varying learning rate, to find the initial learning rate at which training of the proposed architecture is comparatively faster. Then, we compared the proposed architecture with its close variants in terms of classification accuracy, convergence rate and learnable parameters, and finally, we compared the proposed architecture with standard deep learning architectures [9–11] in terms of learnable parameters. Results of the proposed architecture were again compared with results of several state-of-the-art methods of human activity recognition.

5.1 Analysis of the proposed micro-network in designing of CNN architecture

In this section, firstly we evaluated the impact of the micro-networks (micro-network-A and micro-network-B) given in Fig. 2 on the proposed CNN architecture. For this, we designed CNN architectures by separately stacking the micro-network-A and micro-network-B. The architectural details of the designed architectures with micro-network-A and micro-network-B are given in Table 2.

Both the architectures were trained from scratch on YouTube dataset [35] up to 100 epochs with learning rate 0.05 and SGD optimizer. The results of both the networks are compared in terms of convergence rate and learning parameters. The comparative results of are given in Table 3.

From Table 3, it can be observed that the CNN architecture designed with micro-network-B is better than micro-network-A in terms of speed and computational resources. CNN architecture with micro-network-B achieves faster convergence rate and requires lesser number of learnable parameters than the architecture designed with micro-network-A. Therefore, we used micro-network-B (Fig. 2b) to design the proposed architecture for human activity recognition.

After that, we studied the impact of micro-networks in the layers of CNN architecture instead of simply stacking the convolutional layer. For this, we have taken eight different close variants of the proposed CNN architecture by simply stacking convolutional layers instead of using micro-networks. The first architecture (CNN-1) has same depth as the proposed network with convolutional kernel of size 3×3. The second, third and fourth architectures (CNN-2, CNN-3 and CNN-4) have more depth than the proposed network with convolutional kernel of size 3×3. The fifth architecture (CNN-5) has the same depth as proposed network with convolutional kernel of size 5×5. The sixth, seventh and eighth architectures (CNN-6, CNN-7 and CNN-8)) have same width but more depth than the proposed network with convolutional kernel of size 5×5. The architectural details of these architectures are presented in Table 4.

The proposed architecture and all architectures listed in Table 4 are trained up to 50 epochs from scratch on YouTube dataset [35] with learning rate 0.05 and by using SGD optimizer. The experimental results of the proposed architecture and architectures mentioned in Table 4 are given in Table 5 in terms of classification accuracy and learning parameters.

From Table 5, it can be observe that the proposed method achieves highest accuracy value with comparatively less number of learning parameters at 50 epochs. Table 5 shows outperformance of the proposed method over other CNN architectures designed by naively stacking convolutional kernels. The experimental results have demonstrated that by employing varying field size convolutional kernels in the layers of the proposed CNN architecture instead of simply stacking convolutional kernels, we can encapsulate the dynamic range of complex patterns of human activities, which captures information that is distributed globally and locally and helps in faster learning of network from the complex pattern of visual data. The micro-network based architecture is found better in terms of convergence rate, computational resources and classification accuracy. Therefore, for entire experimentation, we have used micro-network-B (as shown in Fig. 2b) for designing of the proposed CNN architecture for human activity recognition.

5.2 Evaluation of the proposed deep CNN architecture

The obtained results using the proposed architecture were evaluated in terms of classification accuracy, and learnable parameters for which, firstly, we compared the proposed architecture with its close variants (Variant_1 and Variant_2). The architectural details of the proposed architecture and its close variants are given in Table 6. The variants of micro-network architectures are shown in Fig. 5a and b.

In variant_1 first two micro-networks consist of 64, (64, 96), (8, 16) , 16 convolutional kernels, next two micro-networks have 96, (96,128), (16, 32), 32 convolutional kernels, and last two have 128, (128, 256), (32, 64), 64 convolutional kernels, whereas in variant_2, first two micro-networks consist of 64, (64, 96), (8, 8, 16, 16) convolutional kernels, the next two micro-networks have

Table 2 Architectural details of the CNN architecture designed with micro-network-A and micro-network-B

Layers	The CNN architecture with micro-network-A	The CNN architecture with micro-network-B
Convolution	32, Conv, stride 2, ReLU, BN	32, Conv, stride 2, ReLU, BN
Pooling	Max pooling, stride 2	Max pooling, stride 2
Convolution	32, Conv, stride 1, ReLU BN	32, Conv, stride 1, ReLU
Convolution	64, Conv, stride 1, ReLU, BN	64, Conv, stride 1, ReLU, BN
Pooling	Max pooling, stride 2	Max pooling, stride 2
Micro-network	$\begin{bmatrix} 1\times 1Conv \\ 1\times 1, 3\times 3Conv \\ 1\times 1, 5\times 5Conv \end{bmatrix}\times 2$	$\begin{bmatrix} 1\times 1Conv \\ 1\times 1, 3\times 3Conv \\ 1\times 1, 3\times 3, 3\times 3Conv \end{bmatrix}\times 2$
Pooling	Max pooling, stride 2	Max pooling, stride 2
Micro-network	$\begin{bmatrix} 1\times 1Conv \\ 1\times 1, 3\times 3Conv \\ 1\times 1, 5\times 5Conv \end{bmatrix}\times 2$	$\begin{bmatrix} 1\times 1Conv \\ 1\times 1, 3\times 3Conv \\ 1\times 1, 3\times 3, 3\times 3Conv \end{bmatrix}\times 2$
Pooling	Max pooling, stride 2	Max pooling, stride 2
Micro-network	$\begin{bmatrix} 1\times 1Conv \\ 1\times 1, 3\times 3Conv \\ 1\times 1, 5\times 5Conv \end{bmatrix}\times 2$	$\begin{bmatrix} 1\times 1Conv \\ 1\times 1, 3\times 3Conv \\ 1\times 1, 3\times 3, 3\times 3Conv \end{bmatrix}\times 2$
Pooling	Average pooling, stride 2	Average pooling, stride 2
FC	256, ReLU	256, ReLU
Classification layer	Softmax classifier	Softmax classifier

Table 3 Comparison of the CNN architectures designed with micro-network-A and micro-network-B in terms of convergence rate and learning parameters

Architectures	Convergence rate (%)	Learning parameters
The CNN architecture with micro-network-A	97.12	1,722,619
The CNN architecture with micro-network-B	98.68	1,542,331

96, (96,128), (16, 16, 32, 32) convolutional kernels, and the last two have 128, (128,256), (32, 32, 64, 64) convolutional kernels. The proposed architecture and its close variants are trained from scratch using raw RGB data of HMDB51 dataset [37] and using SGD optimizer at 0.05 initial learning rate with 1000 epochs. The experimental results of the proposed method and its close variants are shown in Table 7.

From Table 7, it can be observed that the proposed architecture achieves better accuracy (96.58%) than its variants at 1000 epochs. Although variant_2 needs fewer learnable parameters than the proposed architecture but have slow convergence than the proposed architecture, the variant_1 has 96.35% accuracy and variant_2 has 95.46% accuracy in 1000 epochs. Therefore, from Table 7, it can be observed that the proposed architecture outperformed among all its variants.

We again experimented the proposed architecture to investigate the impact of varying learning rate on classification accuracy. For this, we experimented on HMDB51 dataset [37]. We trained the proposed architecture from scratch using SGD optimizer up to 1000 epochs to analyze the network. The results at varying learning rate are given in Table 8. The learning rate plays a vital role in convergence of network. From Table 8, it can be observed that we achieved the highest accuracy value 96.% at learning rate 0.01. Smaller learning rate slows down the network convergence, while larger learning rate sometimes skips the local minima of the loss function. Thus, it can be concluded that the proposed architecture have faster convergence at 0.01 learning rate. Therefore, we have taken an initial learning rate 0.01 for the entire experimentation.

Further, we compared the proposed architecture with standard deep learning architectures, i.e., AlexNet [9], VGGNet [10] and GoogleNet [11] in terms of learnable parameters as shown in Table 9.

It can be seen from Table 9 that the proposed architecture requires lesser number of learnable parameters than standard deep architectures. Therefore, the proposed architecture is computationally efficient in terms of computational resources and time. The fewer number of learnable parameters make the proposed architecture less prone to overfitting problem as one can observe in the learning curve of the proposed architecture shown in Fig. 7. These curves indicate that the proposed architecture

Table 4 Architectural details of the CNN architecture by stacking single convolutional operations in layers

CNN-1	CNN-2	CNN-3	CNN-4	CNN-5	CNN-6	CNN-7	CNN-8
3×3,32	3×3,32	3×3,32	3×3,32	5×5,32	5×5,32	5×5,32	5×5,32
Pooling	Pooling	Pooling	Pooling	Pooling	Pooling	Pooling	Pooling
[3×3,32]	[3×3,32]	[3×3,32]	[3×3,32]	[5×5,32]	[5×5,32]	[5×5,32]	[5×5,32]
[3×3,64]	[3×3,64]	[3×3,64]	[3×3,64]	[5×5,64]	[5×5,64]	[5×5,64]	[5×5,64]
Pooling	Pooling	Pooling	Pooling	Pooling	Pooling	Pooling	Pooling
[3×3,64]	[3×3,64]	[3×3,64]	[3×3,64]	[5×5,64]	[5×5,64]	[5×5,64]	[5×5,64]
[3×3,64]	[3×3,64]	[3×3,64]	[3×3,64]	[5×5,64]	[5×5,64]	[5×5,64]	[5×5,64]
Pooling	Pooling	Pooling	Pooling	Pooling	Pooling	Pooling	Pooling
[3×3,96]	[3×3,96]	[3×3,96]	[3×3,96]	[5×5,96]	[5×5,96]	[5×5,96]	[5×5,96]
[3×3,96]	[3×3,96]	[3×3,96]	[3×3,96]	[5×5,96]	[5×5,96]	[5×5,96]	[5×5,96]
Pooling	Pooling	Pooling	Pooling	Pooling	Pooling	Pooling	Pooling
[3×3,128]	[3×3,128]	[3×3,128]	[3×3,128]	[5×5,128]	[5×5,128]	[5×5,128]	[5×5,128]
[3×3,128]	[3×3,128]	[3×3,128]	[3×3,128]	[5×5,128]	[5×5,128]	[5×5,128]	[5×5,128]
Pooling	Pooling	Pooling	Pooling	Pooling	Pooling	Pooling	Pooling
FC(256)	[3×3,256]	[3×3,256]	[3×3,256]	FC(256)	[5×5,256]	[5×5,256]	[5×5,256]
–	[3×3,256]	[3×3,256]	[3×3,256]	–	[5×5,256]	[5×5,256]	[5×5,256]
–	Pooling	Pooling	Pooling	–	Pooling	Pooling	Pooling
–	FC(256)	[3×3,512]	[3×3,512]	–	FC(256)	[5×5,512]	[5×5,512]
–	–	[3×3,512]	[3×3,512]	–	–	[5×5,512]	[5×5,512]
–	–	Pooling	Pooling	–	–	Pooling	Pooling
–	–	FC(512)	[3×3,1024]	–	–	FC(512)	[5×5,1024]
–	–	–	[3×3,1024]	–	–	–	[5×5,1024]
–	–	–	Pooling	–	–	–	Pooling
–	–	–	FC(1024)	–	–	–	FC(1024)

Table 5 Experimental results of the proposed architecture and architectures mentioned in Table 4

Architectures	Classification accuracy (%)	Learning parameters
CNN-1 (3×3)	93.68	1,026,795
CNN-2 (3×3)	96.35	1,649,899
CNN-3 (3×3)	85.61	5,192,939
CNN-4 (3×3)	75.70	20,143,339
CNN-5 (5×5)	94.90	1,913,003
CNN-6 (5×5)	93.85	4,905,615
CNN-7 (5×5)	74.46	13,943,403
CNN-8 (5×5)	12.76	54,059,627
The proposed architecture	98.68	1,542,331

have capability to be trained on low computing mobile vision devices.

5.3 Comparison of the proposed method with other existing state-of-the-art methods

To evaluate the soundness of the proposed architecture, we conducted experiments on six publically available datasets viz. UCF sports [33], IXMAS [34], YouTube [35], TV-HI [36], HMDB51 [37] and UCF101 [38] and compared the result of the proposed method with other existing state-of-the-art methods in terms of classification accuracy and learning parameter.

First, we experimented the proposed method on UCF sports dataset [33]. We trained the proposed architecture from scratch using RGB data with SGD optimizer and 0.01 learning rate. The network was kept on training until the network achieved global minima. The experimental results of the proposed method and other existing state-of-the-art methods [8, 21, 26, 39–44] on UCF sports dataset are given in Table 10.

From Table 10, one can observe that the proposed method results in the highest value of classification

Table 6 Architectural details of the proposed network architecture and its close variants

Layer	The proposed architecture	Variant_1	Varient_2
Convolution	32, Conv, stride 2, ReLU, BN	32, Conv, stride 2, ReLU, BN	32, Conv, stride 2, ReLU, BN
Pooling	Max pooling, stride 2	Max pooling, stride 2	Max pooling, stride 2
Convolution	32, Conv, stride 1, ReLU BN	32, Conv, stride 1, ReLU BN	32, Conv, stride 1, ReLU BN
Convolution	64, Conv, stride 1, ReLU, BN	64, Conv, stride 1, ReLU, BN	64, Conv, stride 1, ReLU, BN
Pooling	Max pooling, stride 2	Max pooling, stride 2	Max pooling, stride 2
Micro-network	$\begin{bmatrix} 1\times 1Conv \\ 1\times 1, 3\times 3Conv \\ 1\times 1, 3\times 3, 3\times 3Conv \end{bmatrix}\times 2$	$\begin{bmatrix} 1\times 1Conv \\ 1\times 1, 3\times 3Conv \\ 1\times 1, 5\times 5Conv \\ 3\times 3max\_pool, 1\times 1Conv \end{bmatrix}\times 2$	$\begin{bmatrix} 1\times 1Conv \\ 1\times 1, 3\times 3Conv \\ 1\times 1, 3\times 3, 1\times 3, 3\times 1Conv \end{bmatrix}\times 2$
Pooling	Max pooling, stride 2	Max pooling, stride 2	Max pooling, stride 2
Micro-network	$\begin{bmatrix} 1\times 1Conv \\ 1\times 1, 3\times 3Conv \\ 1\times 1, 3\times 3, 3\times 3Conv \end{bmatrix}\times 2$	$\begin{bmatrix} 1\times 1Conv \\ 1\times 1, 3\times 3Conv \\ 1\times 1, 5\times 5Conv \\ 3\times 3max\_pool, 1\times 1Conv \end{bmatrix}\times 2$	$\begin{bmatrix} 1\times 1Conv \\ 1\times 1, 3\times 3Conv \\ 1\times 1, 3\times 3, 1\times 3, 3\times 1Conv \end{bmatrix}\times 2$
Pooling	Max pooling, stride 2	Max pooling, stride 2	Max pooling, stride 2
Micro-network	$\begin{bmatrix} 1\times 1Conv \\ 1\times 1, 3\times 3Conv \\ 1\times 1, 3\times 3, 3\times 3Conv \end{bmatrix}\times 2$	$\begin{bmatrix} 1\times 1Conv \\ 1\times 1, 3\times 3Conv \\ 1\times 1, 5\times 5Conv \\ 3\times 3max\_pool, 1\times 1Conv \end{bmatrix}\times 2$	$\begin{bmatrix} 1\times 1Conv \\ 1\times 1, 3\times 3Conv \\ 1\times 1, 3\times 3, 1\times 3, 3\times 1Conv \end{bmatrix}\times 2$
Pooling	Average pooling, stride 2	Average pooling, stride 2	Average pooling, stride 2
FC	256, ReLU	256, ReLU	256, ReLU
Classification layer	Softmax classifier	Softmax classifier	Softmax classifier

Table 7 Comparison of the proposed architecture with its variants in terms of classification accuracy and learnable parameters

Architectures	Classification accuracy (%)	Learning parameters
Variant_1	96.35	1,674,179
Variant_2	95.46	1,480,307
The proposed architecture	96.58	1,552,099

Table 8 Performance of the proposed architecture at varying learning rates in terms of classification accuracy

Learning rate (lr)	Classification accuracy (%)
0.1	96.03
0.01	96.63
0.05	96.58
0.001	95.44
0.005	96.09

Table 9 Comparison of the proposed architecture with standard deep learning architectures in terms of learnable parameters

Architectures	Learnable parameters (Approx)
AlexNet	71,966,795
VGGNet	139,789,939
GoogleNet	7,427,409
The proposed architecture	1,552,099

accuracy (100%) which is comparable to results of the method proposed by Muhammad et al. [43] with second highest accuracy value (99.90%) and Muhammad et al. [21] with the third-highest accuracy value (99.40%). But these methods are computationally costly than the proposed architecture as one can observe from Table 10 that the method proposed by Muhammad et al. [43] is taking more learning parameter, i.e., 37M than the proposed method (1.5M). The integration of deep learning features computed by using AlexNet and conventional handcrafted features, followed by machine learning classification, makes the method proposed by Muhammad et al. [43], computationally expensive than the proposed method. The method of Muhammad et al. [21] have used a weighted segmentation approach followed by rank correlation-based

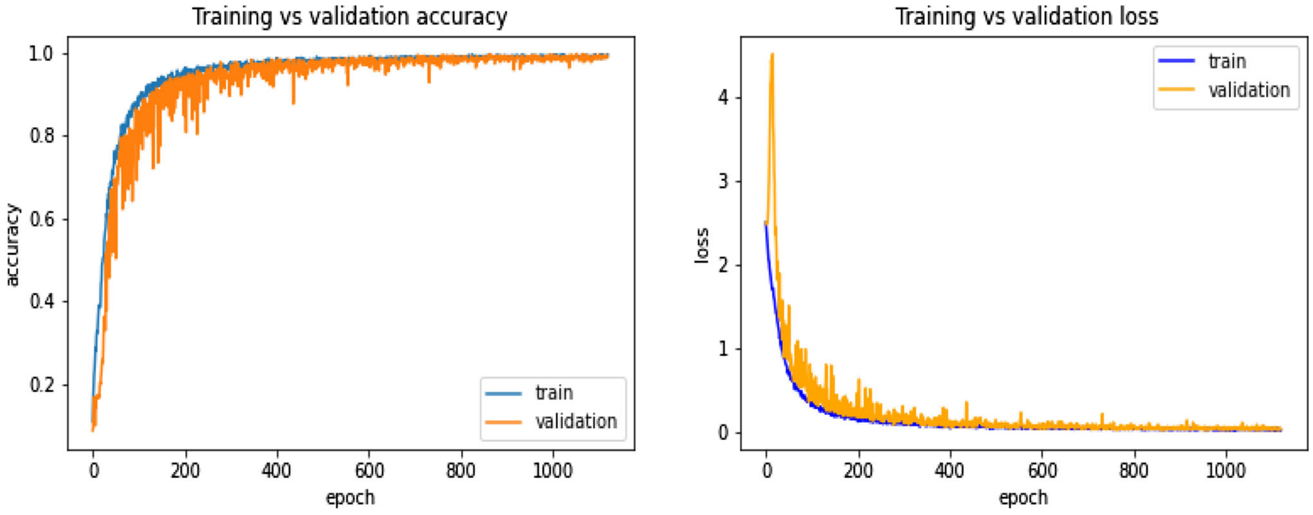


Fig. 7 Learning curves for IXMAS [33] dataset (First column—Training versus validation accuracy and second column—Training versus validation loss)

feature selection approach and then used KNN for classification. Although this method performs well with small size dataset but this technique will become computationally expensive and very difficult to train in case of larger activity category dataset. Therefore, the proposed method perform better than other existing state-of-the-art methods and is computationally efficient in terms of resource and time. The reason behind good performance of the proposed method is the use of varying size receptive field (small micro-network) of the convolutional kernel in different layers of the proposed network. This micro-network with varying receptive field size extracts finer details of several local and global structural information of each activity categories. These extracted information are sufficient to uniquely discriminate each activity categories. The results of this experiment have demonstrated that the proposed method is suitable for small-size realistic dataset that have challenges like large inter-class variations in the same activity category caused due to variations in scale, pose and viewpoint changes. It also works well with low computing devices having limited computational budget.

Second experiment was conducted on the IXMAS dataset [34] which is a low-resolution multi-view dataset. We trained the proposed architecture from scratch using raw RGB data, SGD optimizer with initial learning 0.01 and achieved 99.51% classification accuracy. The experimental result of the proposed method on the IXMAS dataset [34] and its comparison with other existing state-of-the-art methods [18, 26, 43, 45–47] are presented in Table 11.

From Table 11, it can be observed that the proposed method achieved the second-highest classification accuracy, i.e., 99.51%. Although Muhammad et al. [43] have achieved the highest accuracy value, i.e., 99.60% and method of Arati et al. [18] have achieved 99.50% classification accuracy, which are comparable to the performance of the proposed work, but these methods are computationally expensive than the proposed method. Muhammad et al. [43] have used integration of deep

Table 10 Classification accuracy and learning parameters on the UCF sports dataset [33]

Methods	Classification accuracy (%)	Learning parameters (Approx)
Muhammad et al. [8]	99.20	170M
Muhammad et al. [21]	99.40	–
Muhammad et al. [26]	98.00	139M
Amin et al. [39]	82.14	1030M
Amany et al. [40]	86.70	72K
Farhat et al. [41]	99.30	–
Amany al. [42]	92.67	2M
Muhammad et al. [43]	99.90	37M
Saima et al. [44]	97.30	–
The proposed method	100.00	1.5M

learning feature (computed using VGGNet) and handcrafted features followed by use of multiclass support vector machine for activity recognition. This makes the method proposed by Muhammad et al. [43] computationally expensive than the proposed architecture in terms of computational time and resources, as it requires 37M learning parameters. Further, the method of Arati et al. [18] is based on conventional handcrafted feature-based approach in which they used integration of multiple features to extract feature vector followed by multiclass classification. Although this approach requires less computational resources than the deep learning-based methods but it is not suitable for large size datasets having fuzzy boundaries between different activity categories. Thus, the proposed method is found computationally efficient and gives comparable performance to other existing state-of-the-art methods [18, 26, 43, 45–47] in terms of computational cost. The experiment on IXMAS dataset also indicate that the proposed method is also found suitable for low-resolution and multi-view camera environments.

Further, we conducted experiment on YouTube dataset [35] which is realistic sports video dataset. The proposed architecture was trained from scratch, using raw RGB data, SGD optimizer and learning rate 0.01. The proposed network get optimized at 201 epochs and gives 99.70% accuracy value. The experimental results of the proposed method, on YouTube dataset [35], have been compared with other existing state-of-the-art methods [25, 26, 41–44, 48] in terms of classification accuracy value and learning parameters. The results are given in Table 12.

Table 12 shows that the proposed method performs comparatively better than other state-of-the-art methods in terms of classification accuracy and learnable parameter both. The proposed architecture achieves second highest classification accuracy, i.e., 99.70% and requires 1.5 in learnable parameters. Although method proposed by Muhammad et al. [43] achieved the highest accuracy value, i.e., 100%, it is computationally expensive as it can be observed from Table 12 that it requires 37M parameters. Thus, the proposed method has been found computationally efficient and gives comparable performance to other existing state-of-the-art methods [25, 26, 41–44, 48] in terms of computational cost and classification accuracy.

The next experiment was conducted on TV-HI dataset [36] which is on unconstrained realistic video dataset taken from 20 TV shows. The proposed architecture was trained from scratch, with raw RGB data, SGD optimizer and learning rate 0.01. The network achieved global optima at 1000 epochs. The experimental results of the proposed method, on TV-HI dataset [45], have been compared with other existing state-of-the-art methods [49–54] in terms of classification accuracy and learning parameters. The comparison results are given in Table 13.

From Table 13, it can be observed that the proposed method outperformed over other existing state-of-the-art methods in terms of classification accuracy and computational resources both. The proposed method achieved 99.71% classification accuracy value.

This set of experiment also suggest that the proposed method is found suitable for smaller-size realistic dataset, as TV-HI dataset has only four activity classes and that too without any overfitting problem (see Learning curve of this experiment in Fig. 8). The reason behind is that the proposed method extracts sufficient finer details of complex human activities to represent each activity category uniquely and deal with challenges like large inter-class variations in the same activity category caused due to variations in scale, pose, and viewpoint changes and complex motions of human objects.

To validate the effectiveness of the proposed architecture, further, we experimented by using HMDB51 dataset [37] which is one of the most challenging publically available realistic dataset having a total of 51 activity categories. We trained the proposed architecture from scratch by using the SGD optimizer with a learning rate 0.01 and trained the network till it achieved global minima. The proposed method is compared with other state-of-the art methods in terms of classification accuracy and learning parameters. The experimental results of the proposed method and the other existing state-of-the-art methods [8, 24–28, 48, 55, 56] on HMDB51 dataset [37] are given in Table 14.

Table 11 Classification accuracy and learning parameters on the IXMAS dataset [34]

Methods	Classification accuracy (%)	Learning parameters (Approx)
Arati et al. [18]	99.50	–
Muhammad et al. [26]	95.20	139M
Mariem et al. [45]	92.18	–
An-An et al. [46]	94.70	–
Zan et al. [47]	95.10	–
Muhammad et al. [43]	99.60	37M
The proposed method	99.51	1.5M

Table 12 Classification accuracy and learning parameters on YouTube dataset [35]

Methods	Classification accuracy (%)	Learning parameters (Approx)
Noor et al. [25]	97.65	23.5M
Muhammad et al. [26]	99.40	139M
Farhat et al. [41]	94.50	–
Amany et al. [42]	81.40	2M
Muhammad et al. [43]	100	37M
Saima et al. [44]	96.70	–
Zufan et al. [48]	98.20	135M
The proposed method	99.70	1.5M

Table 13 Classification accuracy and learning parameters on TV-HI dataset [36]

Methods	Classification accuracy (%)	Learning parameters (Approx)
Mahlagha et al. [49]	68.00	3M
Mahlagha et al. [50]	75.00	140M
Umair et al. [51]	84.00	44M
Ke et al. [52]	64.00	12M
Qiuhong et al. [53]	55.40	–
Hanli et al. [54]	78.20	–
The proposed method	99.71	1.5M

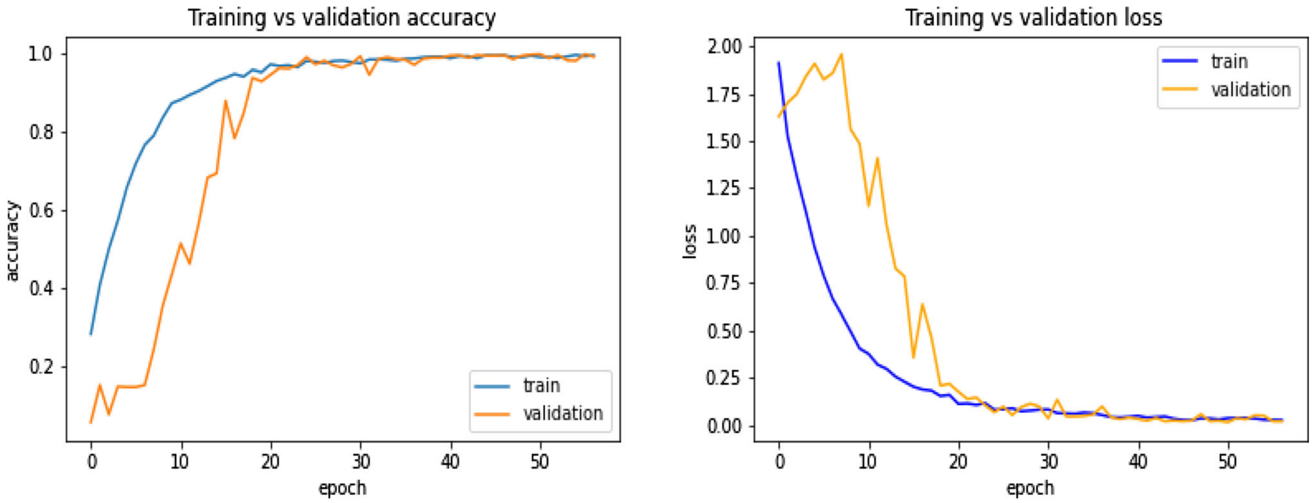


Fig. 8 Learning curves for TV-HI [36] dataset (First column—Training versus validation accuracy and second column—Training versus validation loss)

Table 14 shows that the proposed method has achieved the highest accuracy value (97.48%) and is computationally efficient as it requires only 1.5M trainable parameters. This is due to the multiscale processing capability of the proposed deep architecture, which extract unique discriminative information from realistic scenes that have challenges like extremely large variations in size, orientation and pose of human objects in the same activity category and fuzzy boundary between the activity categories. This experiment also demonstrate that the proposed architecture is also found suitable for realistic larger-size video data and is computationally efficient.

At last, we evaluated the effectiveness of the proposed method and conducted experiment on UCF101 dataset [38], which is one of the most challenging large size dataset having 101 activity categories. We trained the proposed architecture from scratch with SGD optimizer and an initial learning rate 0.01. We trained the network until it achieved global minima. To analyze the effectiveness of the proposed architecture, we compared the results

Table 14 Classification accuracy value and learning parameters on the HMDB51 dataset [37]

Methods	Classification accuracy (%)	Learning parameters (Approx)
Muhammad et al. [8]	81.40	170M
Hao et al. [24]	65.40	57.8M
Noor et al. [25]	95.04	23.5M
Muhammad et al. [26]	93.70	139M
Du et al. [27]	78.70	33.3M
Sachin et al. [28]	96.03	138M
Zufan et al. [48]	60.40	135M
Jun et al. [55]	80.50	47.6M
Sheng et al. [56]	68.20	27.6M
Chaolong et al. [57]	70.80	210M
The proposed method	97.48	1.5M

Table 15 Classification accuracy value and learning parameters on the UCF101 [38] dataset

Methods	Classification accuracy (%)	Learning parameters (Approx)
Hao et al. [24]	92.60	57.8M
Noor et al. [25]	98.66	23.5M
Du et al. [27]	97.30	33.3M
Sachin et al. [28]	97.70	138M
Mei et al. [29]	89.00	60.5M
Zufan et al. [48]	91.00	135M
Jun et al. [55]	98.50	47.6M
Sheng et al. [56]	91.40	27.6M
Chaolong et al. [57]	95.10	210M
Yamin et al. [58]	75.38	44M
The proposed method	98.01	1.5M

of the proposed architecture with several existing state-of-the-art methods [24, 25, 27–29, 48, 55, 56, 58] on classification accuracy and learning parameters. The results are given in Table 15.

From Table 15, it can be observed that the proposed method had achieved the second-highest accuracy value (98.01%). Although the method proposed by Noor et al. [25] achieved the highest classification accuracy value (98.66%), this method is computationally expensive, as requires 23.5M learning parameters, whereas the proposed method requires only 1.5M parameters for learning. Method proposed by Noor et al. [25] firstly used video summarization as a preprocessing step and then hand-crafted feature computation is done followed by use of 3D CNN architecture for activity recognition, which requires more computation time and resources. From this experiment, we also found that the proposed method is flexible to smaller as well as larger size realistic and multi-view datasets and that too within the limited computational budget.

From the close observations of the several experimentations performed over challenging datasets [33–38] and their results shown in Tables 3, 5, 7, 8, 9, 10, 11, 12, 13, 14 and 15, it can be concluded that the proposed architecture is computationally efficient and gives comparatively improved classification results in terms of classification accuracy, learning parameters and convergence rate. It can also be observed from Table 9 that the proposed architecture requires fewer number of learnable parameters than the standard deep architectures [9–11]. Therefore, it requires lesser computational resources, which clearly indicate that the proposed architecture is suitable for low computing devices. The learning curves as shown in Figs. 7 and 8, respectively, demonstrated that the proposed architecture also perform well with smaller size datasets and is less prone to overfitting problem. Again, from Tables 10, 11, 12, 13, 14 and 15, it can be observed that the proposed architecture perform better than other existing state-of-the-art methods in terms of computational efficiency and classification accuracy. Thus, we found that the proposed architecture is computationally efficient in terms of computation time and computational resources. It has also been found that the proposed method is suitable for both larger and smaller size datasets with realistic complex scenarios.

6 Conclusions

In this study, we presented a simple and computationally efficient deep CNN architecture based on the concept of multiscale processing, for human activity recognition in realistic and multi-view environment. In this work, firstly, we designed small micro-networks with varying receptive field size (1×1, 3×3, 5×5). By carefully stacking the varying receptive field size micro-network, a simple and computationally efficient deep CNN architecture has been designed. The designed architecture have potential to capture the dynamic range of complex visual patterns of activity data having challenges of large inter and intra-class variations, fuzzy boundary between activity categories, variations in views, object size, pose, and orientation of human objects in the sample frames of same activity category. The proposed architecture is simple and uses much smaller number of parameters than the standard deep CNN architectures. Use of lesser number of parameters results in better utilization of resources. The proposed architecture is fine-tuned and trained from scratch over several publically available datasets [33–38] and evaluated in terms of convergence rate, classification accuracy and learning parameters. The results and their comparisons with other state-of-the-art methods demonstrate the supremacy of the proposed architecture. Architecture of the proposed method can be trained with any size dataset. From the experiments, results and its exhaustive analyses (from Tables 10, 11, 12, 13, 14 and 15 and Figs. 7 and 8), it has been shown that the proposed architecture is suitable for realistic and multi-view video data that have range of challenges. As a future work, the proposed architecture may be further optimized in terms of learning parameters and increased in depth of the network with limited computational budget. Further, the proposed architecture can be investigated for other computer vision applications.

Acknowledgements This work was supported by the Science and Engineering Research Board (SERB), Department of Science and Technology (DST), New Delhi, India, under Grant No. CRG/2020/001982.

Data availability The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

Conflict of interest There is no conflict of interest.

References

1. Ke Shian-Ru, Le Uyen Hoang, Thuc Yong-Jin Lee, Hwang Jenq-Neng, Yoo Jang-Hee, Choi Kyoung-Ho (2013) A review on video-based human activity recognition. Computers 2(2):88–131
2. Kushwaha Arati, Khare Ashish, Khare Manish (2022) Human activity recognition algorithm in video sequences based on integration of magnitude and orientation information of optical flow. Int J Image Gr 22(01):2250009
3. Dalal Navneet, Triggs Bill (2005) Histograms of oriented gradients for human detection. In: 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05), vol 1, pp 886–893. IEEE
4. Srivastava Prashant, Khare Ashish (2018) Utilizing multiscale local binary pattern for content-based image retrieval. Multimed Tools Appl 77(10):12377–12403
5. Tan Xiaoyang, Triggs Bill (2010) Enhanced local texture feature sets for face recognition under difficult lighting conditions. IEEE Trans Image Process 19(6):1635–1650
6. Laptev Ivan (2005) On space-time interest points. Int J Comput Vision 64(2):107–123
7. Sipiran Ivan, Bustos Benjamin (2011) Harris 3d: a robust extension of the harris operator for interest point detection on 3d meshes. Vis Comput 27(11):963–976
8. Khan Muhammad Attique, Zhang Yu-Dong, Khan Sajid Ali, Attique Muhammad, Rehman Amjad, Seo Sanghyun (2021) A resource conscious human action recognition framework using 26-layered deep convolutional neural network. Multimed Tools Appl 80(28):35827–35849
9. Krizhevsky Alex, Sutskever Ilya, Hinton Geoffrey E (2012) Imagenet classification with deep convolutional neural networks. Adv Neural Inf Process Syst 25:1–9
10. Simonyan Karenl, Zisserman Andrew (2014) Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, pp 1–14
11. Szegedy Christian, Liu Wei, Jia Yangqing, Sermanet Pierre, Reed Scott, Anguelov Dragomir, Erhan Dumitru, Vanhoucke Vincent, Rabinovich Andrew (2015) Going deeper with convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1–9
12. Khare Manish, Srivastava Rajneesh Kumar, Khare Ashish (2014) Single change detection-based moving object segmentation by using daubechies complex wavelet transform. IET Image Proc 8(6):334–344
13. Srivastava Yash, Murali Vaishnav, Dubey Shiv Ram (2019) A performance evaluation of loss functions for deep face recognition. In: National conference on computer vision, pattern recognition, image processing, and graphics, pp 322–332. Springer
14. Hsu Pai-Hui, Zhuang Zong-Yi (2020) Incorporating handcrafted features into deep learning for point cloud classification. Remote Sens 12(22):3713
15. Nadjet Bouchaour, Smaine Mazouzi (2022) Deep pattern-based tumor segmentation in brain mris. Neural Comput Appl 34(17):14317–14326
16. Yang Ziheng, Benhabiles Halim, Hammoudi Karim, Windal Feryal, He Ruiwen, Collard Dominique (2021) A generalized deep learning-based framework for assistance to the human malaria diagnosis from microscopic images. Neural Computing and Applications, pp 1-16
17. Roitberg Alina, Perzylo Alexander, Somani Nikhil, Giuliani Manuel, Rickert Markus, Knoll Alois (2014) Human activity recognition in the context of industrial human-robot interaction. In: signal and information processing association annual summit and conference (APSIPA), 2014 Asia-Pacific, pp 1–10. IEEE

18. Kushwaha Arati, Khare Ashish, Srivastava Prashant (2021) On integration of multiple features for human activity recognition in video sequences. Multimed Tools Appl 80(21):32511–32538
19. Singh Roshan, Kushwaha Alok Kumar Singh, Srivastava Rajeev (2019) Multi-view recognition system for human activity based on multiple features for video surveillance system. Multimed Tools Appl 78(12):17165–17196
20. Nigam Swati, Khare Ashish (2016) Integration of moment invariants and uniform local binary patterns for human activity recognition in video sequences. Multimed Tools Appl 75(24):17303–17332
21. Sharif Muhammad, Khan Muhammad Attique, Zahid Farooq, Shah Jamal Hussain, Akram Tallha (2020) Human action recognition: a framework of statistical weighted segmentation and rank correlation-based selection. Pattern Anal Appl 23(1):281–294
22. Xiao Guoqing, Li Jingning, Chen Yuedan, Li Kenli (2020) Malfcs: an effective malware classification framework with automated feature extraction based on deep convolutional neural networks. J Parallel Distrib Comput 141:49–58
23. Xiao G, Li K, Zhou X, Li K (2017) Efficient monochromatic and bichromatic probabilistic reverse top-k query processing for uncertain big data. J Comput Syst Sci 89:92–113
24. Yang Hao, Yuan Chunfeng, Li Bing, Yang Du, Xing Junliang, Weiming Hu, Maybank Stephen J (2019) Asymmetric 3d convolutional neural networks for action recognition. Pattern Recogn 85:1–12
25. Almaadeed Noor, Elharrouss Omar, Al-Maadeed Somaya, Bouridane Ahmed, Beghdadi Azeddine (2019) A novel approach for robust multi human action recognition and summarization based on 3d convolutional neural networks. arXiv preprint arXiv:1907.11272, pp 1–22
26. Khan Muhammad Attique, Javed Kashif, Khan Sajid Ali, Saba Tanzila, Habib Usman, Khan Junaid Ali, Abbasi Aaqif Afzaal (2020) Human action recognition using fusion of multiview and deep features: an application to video surveillance. Multimedia tools and applications, pp 1–27
27. Tran Du, Wang Heng, Torresani Lorenzo, Ray Jamie, LeCun Yann, Paluri Manohar (2018) A closer look at spatiotemporal convolutions for action recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 6450–6459
28. Chaudhary Sachin, Murala Subrahmanyam (2019) Depth-based end-to-end deep network for human action recognition. IET Comput Vis 13(1):15–22
29. Leong Mei Chee, Prasad Dilip K, Lee Yong Tsui, Lin Feng (2020) Semi-cnn architecture for effective spatio-temporal learning in action recognition. Appl Sci 10(2):557
30. Luo Wenjie, Li Yujia, Urtasun Raquel, Zemel Richard (2016) Understanding the effective receptive field in deep convolutional neural networks. Adv Neural Inf Process Syst 29:4905–4913
31. Yamashita Rikiya, Nishio Mizuho, Do Richard Kinh Gian, Togashi Kaori (2018) Convolutional neural networks: an overview and application in radiology. Insights Imaging 9(4):611–629
32. Bottou Léon (2010) Large-scale machine learning with stochastic gradient descent. In: Proceedings of COMPSTAT'2010, pp 177–186. Springer
33. Rodriguez Mikel D, Ahmed Javed, Shah Mubarak (2008) Action mach a spatio-temporal maximum average correlation height filter for action recognition. In: 2008 IEEE conference on computer vision and pattern recognition, pp 1–8. IEEE
34. Kim Sun Jung, Kim Soo Wan, Sandhan Tushar, Choi Jin Young (2014) View invariant action recognition using generalized 4d features. Pattern Recogn Lett 49:40–47
35. Liu Jingen, Luo Jiebo, Shah Mubarak (2009) Recognizing realistic actions from videos "in the wild". In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp 1996–2003. IEEE
36. Patron-Perez Alonso, Marszalek Marcin, Reid Ian, Zisserman Andrew (2012) Structured learning of human interactions in tv shows. IEEE Trans Pattern Anal Mach Intell 34(12):2441–2453
37. Kuehne Hildegard, Jhuang Hueihan, Garrote Estíbaliz, Poggio Tomaso, Serre Thomas (2011) Hmdb: a large video database for human motion recognition. In: 2011 International conference on computer vision, pp 2556–2563. IEEE
38. Soomro Khurram, Zamir Amir Roshan, Shah Mubarak (2012) A dataset of 101 human action classes from videos in the wild. Center for Research in Computer Vision, 2(11)
39. Zare Amin, Moghaddam Hamid Abrishami, Sharifi Arash (2020) Video spatiotemporal mapping for human action recognition by convolutional neural network. Pattern Anal Appl 23(1):265–279
40. Abdelbaky Amany, Aly Saleh (2021) Two-stream spatiotemporal feature fusion for human action recognition. Vis Comput 37(7):1821–1835
41. Afza Farhat, Khan Muhammad Attique, Sharif Muhammad, Kadry Seifedine, Manogaran Gunasekaran, Saba Tanzila, Ashraf Imran, Damaševičius Robertas (2021) A framework of human action recognition using length control features fusion and weighted entropy-variances based feature selection. Image Vis Comput 106:104090
42. Abdelbaky Amany, Aly Saleh (2021) Human action recognition using three orthogonal planes with unsupervised deep convolutional neural network. Multimed Tools Appl 80(13):20019–20043
43. Khan Muhammad Attique, Sharif Muhammad, Akram Tallha, Raza Mudassar, Saba Tanzila, Rehman Amjad (2020) Hand-crafted and deep convolutional neural network features fusion and selection strategy: an application to intelligent human action recognition. Appl Soft Comput 87:105986
44. Nazir Saima, Yousaf Muhammad Haroon, Nebel Jean-Christophe, Velastin Sergio A (2018) A bag of expression framework for improved human action recognition. Pattern Recogn Lett 103:39–45
45. Gnouma Mariem, Ladjailia Ammar, Ejbali Ridha, Zaied Mourad (2019) Stacked sparse autoencoder and history of binary motion image for human activity recognition. Multimed Tools Appl 78(2):2157–2179
46. Liu An-An, Yu-Ting Su, Nie Wei-Zhi, Kankanhalli Mohan (2016) Hierarchical clustering multi-task learning for joint human action grouping and recognition. IEEE Trans Pattern Anal Mach Intell 39(1):102–114
47. Gao Zan, Nie Weizhi, Liu Anan, Zhang Hua (2016) Evaluation of local spatial-temporal features for cross-view action recognition. Neurocomputing 173:110–117
48. Zhang Zufan, Lv Zongming, Gan Chenquan, Zhu Qingyi (2020) Human action recognition using convolutional lstm and fully-connected lstm with different attentions. Neurocomputing 410:304–316
49. Afrasiabi Mahlagha, Mansoorizadeh Muharram et al (2020) Dtw-cnn: time series-based human interaction prediction in videos using cnn-extracted features. Vis Comput 36(6):1127–1139
50. Afrasiabi Mahlagha, Khotanlou Hassan, Gevers Theo (2020) Spatial-temporal dual-actor cnn for human interaction prediction in video. Multimed Tools Appl 79(27):20019–20038
51. Haroon Umair, Ullah Amin, Hussain Tanveer, Ullah Waseem, Sajjad Muhammad, Muhammad Khan, Lee Mi Young, Baik Sung Wook (2022) A multi-stream sequence learning framework for human interaction recognition. IEEE Trans Human-Mach Syst 52(3):435–444
52. Ke Qiuhong, Bennamoun Mohammed, An Senjian, Boussaid Farid, Sohel Ferdous (2016) Human interaction prediction using

deep temporal features. In: European conference on computer vision, pp 403–414. Springer

53. Jeongmin Yu, Jeon Moongu, Pedrycz Witold (2014) Weighted feature trajectories and concatenated bag-of-features for action recognition. Neurocomputing 131:200–207
54. Wang Hanli, Yi Yun, Wu Jun (2015) Human action recognition with trajectory based covariance descriptor in unconstrained videos. In: Proceedings of the 23rd ACM international conference on Multimedia, pp 1175–1178
55. Li Jun, Liu Xianglong, Zhang Wenxuan, Zhang Mingyuan, Song Jingkuan, Sebe Nicu (2020) Spatio-temporal attention networks for action recognition and detection. IEEE Trans Multimed 22(11):2990–3001
56. Sheng Yu, Xie Li, Liu Lin, Xia Daoxun (2019) Learning long-term temporal features with deep neural networks for human action recognition. IEEE Access 8:1840–1850
57. Zhang C, Xu Y, Xu Z, Huang J, Lu J (2022) Hybrid handcrafted and learned feature framework for human action recognition. Appl Intell 52(11):12771–12787
58. Han Yamin, Zhang Peng, Zhuo Tao, Huang Wei, Zhang Yanning (2017) Video action recognition based on deeper convolution networks with pair-wise frame motion concatenation. In: Proceedings of the IEEE conference on computer vision and pattern recognition Workshops, pp 8–17

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.