Contents lists available at ScienceDirect

Future Generation Computer Systems

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

Human action recognition using attention based LSTM network with dilated CNN features

Khan Muhammad [a,*], Mustaqeem [b], Amin Ullah [b,c], Ali Shariq Imran [d], Muhammad Sajjad [d,e,**], Mustafa Servet Kiran [f], Giovanna Sannino [g], Victor Hugo C. de Albuquerque [h,i]

[a] *Visual Analytics for Knowledge Laboratory (VIS2KNOW Lab), School of Convergence, College of Computing and Informatics, Sungkyunkwan University, Seoul 03063, Republic of Korea*
[b] *Department of Software, Sejong University, Seoul 05006, Republic of Korea*
[c] *CoRIS Institute, Oregon State University, Corvallis 97331, Oregon, USA*
[d] *Color and Visual Computing Lab, Department of Computer Science, Norwegian University of Science and Technology (NTNU), 2815 Gjøvik, Norway*
[e] *Digital Image Processing Laboratory, Department of Computer Science, Islamia College Peshawar, Peshawar 25000, Pakistan*
[f] *Department of Computer Engineering, Faculty of Engineering and Natural Sciences, Konya Technical University, 42075, Konya, Turkey*
[g] *Institute of High-Performance Computing and Networking, National Research Council of Italy, 80131 Naples, Italy*
[h] *Graduate Program on Teleinformatics Engineering, Federal University of Ceará, Fortaleza, Fortaleza/CE, Brazil*
[i] *Graduate Program on Electrical Engineering, Federal University of Ceará, Fortaleza/CE, Brazil*

ARTICLE INFO

Article history:
Received 2 January 2021
Received in revised form 22 March 2021
Accepted 17 June 2021
Available online 24 June 2021

Keywords:
Artificial intelligence
Action recognition
Attention mechanism
Big data
Dilated convolutional neural network
Deep bi-directional LSTM
Multimedia data security

ABSTRACT

Human action recognition in videos is an active area of research in computer vision and pattern recognition. Nowadays, artificial intelligence (AI) based systems are needed for human-behavior assessment and security purposes. The existing action recognition techniques are mainly using pre-trained weights of different AI architectures for the visual representation of video frames in the training stage, which affect the features' discrepancy determination, such as the distinction between the visual and temporal signs. To address this issue, we propose a bi-directional long short-term memory (BiLSTM) based attention mechanism with a dilated convolutional neural network (DCNN) that selectively focuses on effective features in the input frame to recognize the different human actions in the videos. In this diverse network, we use the DCNN layers to extract the salient discriminative features by using the residual blocks to upgrade the features that keep more information than a shallow layer. Furthermore, we feed these features into a BiLSTM to learn the long-term dependencies, which is followed by the attention mechanism to boost the performance and extract the additional high-level selective action related patterns and cues. We further use the center loss with Softmax to improve the loss function that achieves a higher performance in the video-based action classification. The proposed system is evaluated on three benchmarks, i.e., UCF11, UCF sports, and J-HMDB datasets for which it achieved a recognition rate of 98.3%, 99.1%, and 80.2%, respectively, showing 1%–3% improvement compared to the state-of-the-art (SOTA) methods.

1. Introduction

Video based action recognition is an emerging and challenging area of research in this era particularly for identifying and recognizing actions in a video sequence from a surveillance stream. The action recognition in a video has many applications, such as content-based video retrieval [1], surveillance systems for security and privacy purposes [2], human–computer-interaction, and activity recognition [3]. Nowadays, the digital contents are exponentially growing day-by-day, so effective AI-based intelligent internet of things (IoT) systems [2,4] are needed for surveillance to monitor and identify human actions and activities. The aim of action recognition is to detect and identify people, their behavior, suspicious activities in the videos, and deliver appropriate information to support interactive programs and IoT based applications [5,6]. Action recognition still poses many challenges when it comes to ensuring the security and the safety of the residents, including industrial monitoring, violence detection, person identification, virtual reality, and cloud environments [7–10] due to significant improvements in camera movements, occlusions, complex background, and variations in illumination. The spatial and temporal information play a crucial role in recognizing different human actions in videos. In the last decade, most of

[*] Corresponding author at: Visual Analytics for Knowledge Laboratory (VIS2KNOW Lab), School of Convergence, College of Computing and Informatics, Sungkyunkwan University, Seoul 03063, Republic of Korea.
[**] Corresponding author at: Color and Visual Computing Lab, Department of Computer Science, Norwegian University of Science and Technology (NTNU), 2815 Gjøvik, Norway.
E-mail addresses: khan.muhammad@ieee.org (K. Muhammad), muhammad.sajjad@ntnu.no (M. Sajjad).

https://doi.org/10.1016/j.future.2021.06.045
0167-739X/© 2021 Elsevier B.V. All rights reserved.

the methods used handcrafted features engineering to signify the spatial attributes of dynamic motion for characterizing the corresponding action in the videos. The handcrafted based features method in action recognition is mostly database-oriented and cannot satisfy the universal situation due to the motion style and the complex background clutter. In this regard, the representative motion features and the conventional methods are gradually upgraded from 2D to 3D for capturing accurate information. Such techniques transformed spatial features to 3D spatiotemporal features to simultaneously capture the dynamic information in a sequence of frames [11].

Deep learning is now the most dominant and widely used technique for high-level discriminative salient features learning and making end-to-end systems in video based action and behavior recognition [12]. The existing deep learning approaches for human action recognition (HAR) utilize simple convolutional neural networks (CNNs) strategies in convolution operation to learn the features from video frames by using per-trained models. These convolutional layers extract and learn spatial features to train a model for classification. Comparatively, the traditional CNN models have lower performance than handcrafted features in sequential data [13]. Standard CNN models, such as AlexNet, VGG, and ResNet learn spatial features from a single input image. These models are useful in capturing the spatial information, but they are not very effective for temporal data, which is an important factor to capture motion information for the HAR in a video sequence. For instance, Dai et al. [14] used the long short-term memory (LSTM) for action recognition using features learned through a CNN having spatiotemporal information. The video based high-level HAR techniques require a two-stream approach to design separate modules, which learn spatial and temporal features in video sequences by fusing mechanisms to capture dynamic information in sequential data [15]. Recently, the spatiotemporal issues are handled by employing the recurrent neural networks (RNNs), where the LSTM is designed specifically for long-term video sequence to learn and process the temporal features for HAR in surveillance systems [16]. Currently, most of the researchers developed a two-stream approach for action recognition to combine the temporal and spatial features for joint features training to cover the current challenges and limitations of the HAR.

Based on these facts, the precise recognition of action in the real-world videos is still challenging, lacking information about the motion, style, and background clutter for the proper identification of human actions. The traditional methods failed to address these issues, which is due to problems with handling continuous actions, difficulty in modeling crowded scenes due to occlusion, and sensitivity to noise [17]. Similarly, the current methods for the HAR resolved the sequence learning problem by the RNNs, LSTM, and gated recurrent unit but without focusing on selective information in sequences, which is very important for keeping a connection between the previous and the next frames. To address this problem, we propose a novel attention-based HAR system to learn spatiotemporal features and selectively focus on discriminative cues in long-term sequences for recognizing actions in video frames, which is the most suitable for a surveillance system. In this system, we use the DCNN with residual blocks to upgrade the learned features and the BiLSTM with attention weights, which selectively focus on the effective features in the input frames sequence and recognize action in a video accordingly. In the proposed system, we find numerous preeminence for action recognition and convolution operation in CNN extracts spatial information, which is processed by the BiLSTM to realize the contents to recommend the human actions in a better way. We process every eighth frame from a video to extract the high-level discriminative information and pass it from the attention mechanism that re-adjusts the attention weights for selective cues in the sequence. Due to these features, the proposed method is more suitable for the HAR for the surveillance video streams, as evident from experimentations. The key contributions of the proposed HAR system are summarized as follows:

1. We propose a novel framework that utilizes a convolutional network with residual blocks for upgrading the features. The skips connection strategy is incorporated into the traditional methods for a better representation of the human actions in the surveillance videos for security purposes.
2. We propose an attention mechanism with a deep BiLSTM network to learn the spatiotemporal features in sequential data. We re-adjust the attention weights with the learned global features to easily detect and focus on significant cues to recognize the human actions in a sequence.
3. We fine-tune and introduce a center loss function to improve the performance of the HAR in the surveillance videos. The center loss function defines the distance between the centers of each feature with a consistent class center. Through extensive experiments, we proved that the center loss with the softmax loss improves the performance of the proposed system compared to the existing methods, validating its significance for the HAR.
4. The proposed HAR system is evaluated over three standard UCF11, UCF Sports, and JHMBD action datasets during experimentations, resulting a high recognition performance of 98.3%, 99.1%, and 80.2%, respectively. Our system secures a high performance over SOTA methods, indicating its applicability and feasibility for deployment in surveillance applications.

The remaining article is structured as follows Sections 2 and 3 illustrate the related literature and the proposed HAR architecture along with its related components, respectively. Section 4 presents the experimentations and results, and the conclusions are drawn along with the possible directions for future work in Section 5.

2. Literature study

Action recognition is a major field of research in the computer vision domain, where researches have introduced numerous techniques [18], utilizing traditional machine learning, AI, and convolutional networks to recognize the human actions in video streams. In the past decade, researchers mostly used traditional machine learning techniques to develop efficient systems for the HAR by features engineering. Nowadays, researchers use deep learning approaches to extract both the sequential and spatial information from a sequence of frames for the actions classification. Hence, an efficient classification method of an action recognition can be used in cybercrimes investigation and an IoT-based security system for intelligent portable devices [19]. A literature review of the related existing methods is given in the subsequent sections.

2.1. Traditional machine learning and handcrafted features-based action recognition

Traditional machine learning based action recognition techniques use three main phases: (i) features extraction using handcrafted feature descriptors, (ii) features representation encoded by an algorithm, (iii) and finally classification of the features with a suitable machine learning algorithm [3]. In computer vision, mostly two types of features extraction techniques are used: local features-based, and global feature-based approaches. In the local features-based approaches, the features are described as independent patches, interest points, and gesture information that match the learned cues for a specific task. In contrast, the global features are represented by the area of interest, which are described by background subtraction and tracking [20,21]. In conventional machine learning, researchers have developed efficient systems for action recognition by utilizing handcrafted features, such as VLAD [15] and BOW [22]. Most handcrafted features extractors are designed for specific datasets, and they are domain specific. They cannot be used for general-purpose features learning [23]. Some researchers used key frame-based strategies for reducing the processing time of their systems [24]. For instance, Yasin et al. [25] developed a key technique for action recognition in video sequence using key frame selection and then utilized it for the HAR. Similarly, Zhao et al. [26] presented a multi-features fusion based HAR using key frames by the conventional machine learning method. The conventional machine learning algorithms have achieved great success in the past decade. However, they are limited by human cognition and still have some problems, such as being time consuming, being labor-intensive, and the selection of features engineering is a cumbersome process [27]. Due to the limitations and challenges in handcrafted based HAR, the researchers moved towards deep learning to design efficient and new techniques for advanced video based HAR systems.

2.2. Deep learning based action recognition

Unlike a three-step traditional machine learning architecture, deep learning introduces a modern end-to-end architecture to learn and represent high-level discriminative visual features and classifies them simultaneously [28]. Popular end-to-end CNN architectures frequently adjust the parameters according to the data and learn the best features using convolutional operations [29]. Typically, the CNN-based features learning approaches used for 2D data to represent the visual information are not suitable for 3D data representations. For instance, some researchers introduced methods to learn features from video frames using 3D filters instead of 2D [30,31]. Their models outperformed the results in the video analysis (action recognition, object tracking, and video retrieval) as compared to 2D CNNs and handcrafted features-based methods. Similarly, Simonyan et al. [32] proposed a two-stream architecture for video action recognition by utilizing a CNN to overcome the problem of extracting the motion cues among the repeated video frames. Feichtenhofer et al. [33] proposed a method to recognize action based on the spatial and motion features. They evaluated the system on the UCF101 to prove the importance of temporal information. Tu et al. [34] used a multi-stream CNN model to learn human related regions that recognized multiple actions in a video. The authors in [14,35] utilized a two-stream LSTM based fused network to learn the spatiotemporal cues in video sequences. Similarly, the researchers developed hybrid techniques [36] to learn multiple features for action recognition by utilizing the features' fusion [37]. Guimaraes et al. [38] developed a technique for anomaly detection by using an optimum-path forest classifier that monitors and detects an anomaly through an IoT-based intelligent system. However, these deep learning approaches are presented to learn and recognize the short-term temporal information, and they are not suitable and sufficient for long-term sequences. With the success of RNN, its advanced version is introduced called LSTM that encodes the long-term dependencies [16]. Nowadays, mostly LSTM networks are used [16] to classify the long sequential data in different domains, such as action recognition, speech processing, and weather prediction. Hence, the redundancy is also problematic for big data so Xu et al. [39] developed a deep learning-based method to avoid the redundancy of big data as well as utilized the reinforcement learning for resource allocation based on the IoT content-centric [40].

2.3. Attention based mechanism

The attention mechanism is recently introduced that achieved promising success in numerous challenging temporal tasks like video captioning, action recognition, and sequence learning. In the action video contents such as walking, running, jogging, etc., a sequence of frames plays an important role at different stages to show diverse attention to the viewers first look [41]. In video and image captioning applications, attention mechanisms reached a SOTA performance on standard datasets to learn and describe the contents. For instance, Bahdanau et al. [42] developed a technique for English to French translation using an attention-based model, which leads the SOTA phrase-based model. In [43], the authors developed a method of extracting informative frames in a video to recognize action by utilizing attention mechanisms. Some researchers presented a unique attention mechanism for action recognition. They designed separate models for learning spatial and temporal information [44,45]. Nowadays, the usage of the two-stream attention mechanism, which indirectly pays attention to sequential information, is vastly used in several applications [46].

3. Proposed methodology

In this section, we explain the proposed architecture of action recognition and its main components, such as the DCNN, residual blocks, deep BiLSTM, and center loss function. We recognize a human action in video frames with the help of a dilated convolution network that is coupled with upgraded features learning blocks, to improve the learned features and the deep BiLSTM network with an attention mechanism. In the proposed system, we first extract the CNN features from the input data using a DCNN. Secondly, we fuse the upgraded features with the dilated features by utilizing the skip connections. A DCNN scheme is utilized in the skip connection to recognize the hidden cues. The fused output is fed to the deep BiLSTM to learn the sequential information followed by attention mechanism to produce a context vector. Finally, the context vector is passed through a full convolution network. Both the center and the Softmax loss are computed, and an aggregated loss is obtained to reach a final decision for the action recognition. The overall architecture of the proposed system is illustrated in Fig. 1, and a detailed description of the main components is explained in the upcoming sections.

3.1. Dilated convolutional neural networks

In this subsection, we explain the strategy of the DCNN and how it extracts features instead of the conventional CNNs. The DCNN uses more inclusive receptive fields for the features abstraction compared to the conventional CNN. We use the DCNN for semantic segmentation and to hold the implicit information in the mask to obtain a better performance. Most architectures in computer vision add several pooling layers to down-sample the image size and increase the receptive field to expand and up-sample the image size. In traditional CNNs, the image size is often changed because of pooling operation due to which some information might be lost. We developed a DCNN to overcome this issue. We used a CNN with 3 × 3 kernels with stride setting one, max-pooling 2 × 2 with stride setting, and chose the "*valid*" padding strategy at the same time. Moreover, we also used the dropout layer in the dilated network.

As an additional part of the DCNN, the upgrade features learning block (UFLB) is designed for extracting the most discriminative salient action features. Every block consists of one DCNN, one batch normalization (BN), and one Leaky relu layer, which is presented in Fig. 2. In the proposed CNN model, we use three

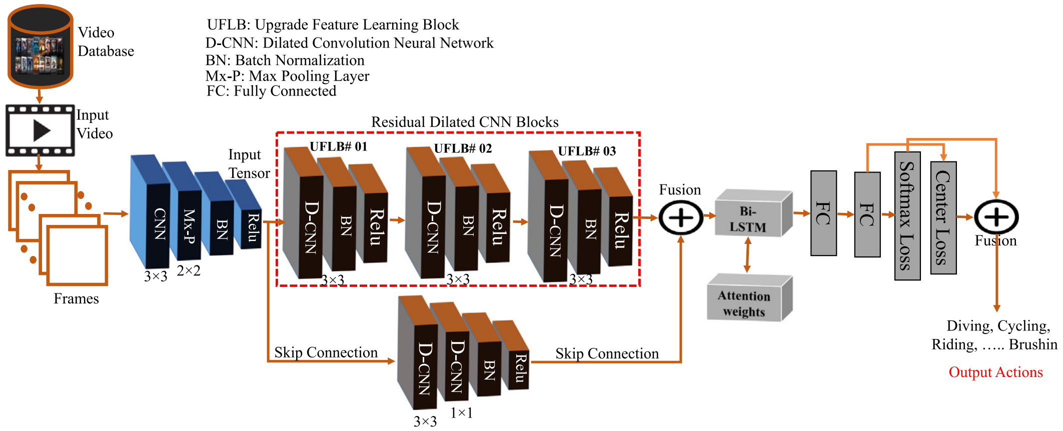

Fig. 1. The complete architecture of the proposed human action recognition system using an attention based BiLSTM with dilated CNN features.

Table 1
The parameters used in the proposed system, where the output dimensions represent the height and width of the feature maps with channels in the UFLBs. The dilation rate is set to 2.

Name	Output dimensions	K-size	Stride
CNN layer	128 × 128 × 128	3 × 3	1 × 1
Max-pooling strategy	64 × 64 × 128	2 × 2	2 × 2
First UFLB	64 × 64 × 128	3 × 3	–
Second UFLB	64 × 64 × 128	3 × 3	–
Third UFLB	64 × 64 × 128	3 × 3	–
BiLSTM	–	256	–
Attention mechanism	–	–	–
Fully connected layer	–	No. actions	–

upgrade learning blocks to learn more prominent features. The BN layer in the DCNN helps normalize the learned features for a better performance with a fast-training speed. It also avoids the vanishing gradient issues in the training process. We utilize the Leaky relu function instead of the relu activation function to define the output of BN layers, as shown in Eq. (1).

$$y_i = \begin{cases} x_i, \text{ if } \quad x_i, \geq 0 \\ \frac{x_i}{a_i}, \text{ if } \quad x_i, < 0 \end{cases} \tag{1}$$

Here "y_i" presents the output of the relu function. In the proposed dilated layers, we utilize the CNN for extracting the local features from the input data and keep the useful information with the help of the dilated convolution layers instead of the pooling strategy. The dilated convolution layer produces the output feature maps from the input data by computing the dot products among the CNN filters and the input. The utilized parameters in the proposed system are illustrated in Table 1.

In the convolution layer, $x(i, j)$ is the input that is convoluted with filter $w(i, j)$ of size $c \times d$ to obtain the results, which are stored in $z(i, j)$, as shown in Eq. (2).

$$z\left(i,j\right) = x\left(i,j\right) \times w\left(i,j\right) = \sum_{a=-c}^{c} \sum_{b=-d}^{d} x\left(a,b\right) \cdot w\left(i-a, j-b\right) \tag{2}$$

The resultant value is subjected to Eq. (3) for the convolution operation followed by Eq. (4) for the BN, which normalizes the input features for a better performance and a fast training process.

$$z_i^l = b_i^l + \sum_j z_i^{l-1} . w_{ij}^l \tag{3}$$

$$z_i^l = \text{ BN}\left(z_i^l\right) = \Upsilon\left(\frac{z_i^l - \mu}{\sqrt{\alpha 2 + \epsilon}}\right) + \beta \tag{4}$$

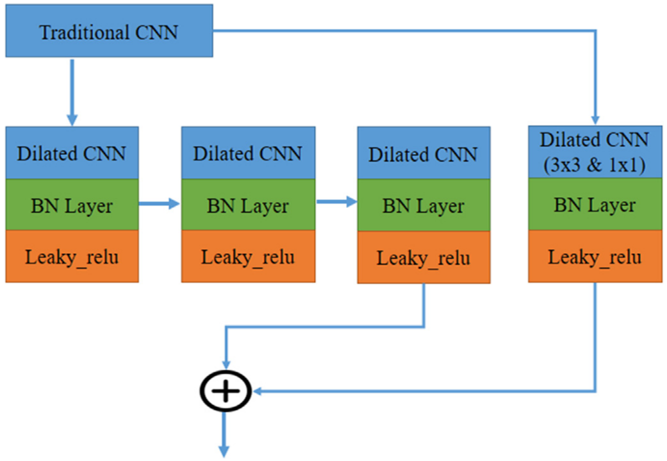

Fig. 2. The upgraded features learning blocks and the skip connections approach with a dilated convolutional network.

Here z_i^l and z_i^{l-1} represent the output and input features in the i^{th} and the j^{th} value at the 1^{st} layer, and the filter among the i^{th} and j^{th} feature value is denoted by w_{ij}^l (kernel) in the convolution. The μ represents the mean, and α^2 represents the variance of the ith output at the 1^{st} layer. We set ϵ and β as the learning rate to show the power of the suggested network.

3.2. Residual block structure

We combine the residual framework with the DCNN at the time of the features learning with the help of a skip connections approach to upgrade the learned features. Eqs. (5) and (6) show the in-depth network layers to further extract and learn the salient high-level features. By using the residual blocks, we utilize not only the current information but also the previous information to train the network after Leaky relu, which is shown in Eqs. (5) and (6). Similarly, we use Eq. (7) to get the output of the residual blocks.

$$z_i^{l-1} = w_i^{l-1} \cdot z_i^{l-2} + b_i^{l-1} \tag{5}$$

$$z_i^{l-1} = \varphi(\text{BN}(z_i^{l-1})) \tag{6}$$

$$z_i^l = \varphi(z_i^l + z_i^{l-2}) \tag{7}$$

In Eqs. (5) and (6), z_i^{l-1} and z_i^{l-2} represent the dilated CNN features in the i^{th} layer, consistent bias, and the weights are indicated by "b" and "w". The symbol "φ" denotes the Leaky ReLU as an activation function of the prescribed network. The resultant fused feature vector of the DCNN and the UFLBs z_i^l is fed into the BiLSTM network for further processing.

3.3. Global feature learning

An RNN is widely used these days that combines the previous frame information in a video frame sequence to the current frame for better action recognition. This model is specially designed for series/continuous data problems. It captures the temporal information in a sequence, which is not possible through the traditional CNN models, such as AlexNet and VGG. RNNs mostly suffer from the "*vanishing gradient*" problem, which can be addressed using the LSTM [47]. The LSTM network includes three gates: input, output, and forget gates. The last forget gate is used for discarding and deciding to remove the irrelevant information in the input x_t and from the previous output h_{t-1} [48]. We assume the discriminative features vector from the CNN model at unit time t, which is fed into the LSTM network via the input, output, and forget gated mechanism with a self-recurrent connection [49, 50] to update the LSTM units. The LSTM updating mechanism is mathematically given in Eqs. (8) to (12).

$$i_t = \sigma(w_i z_i^l + U_i h_{t-1} + b_i) \tag{8}$$

$$f_t = \sigma(w_f z_i^l + U_f h_{t-1} + b_f) \tag{9}$$

$$o_t = \sigma(w_o z_i^l + U_o h_{t-1} + b_0) \tag{10}$$

$$c_t = c_{t-1} \cdot f_t + \tanh(w_c z_i^l + U_c h_{t-1} + b_c) \tag{11}$$

$$z_i^l = \tanh{(c_t)} \cdot o_t \tag{12}$$

Here i_t, f_t, and o_t are the input, forget, and output gates at time $_t$, respectively. The sigmoid function is represented by σ, and the weight matrices are denoted by "w_i", "u_i", "w_f", "u_f", "w_o", and "u_o", respectively with bias "b". In the proposed architecture, we use the BiLSTM network, which not only maintains the history from previous layers but also utilizes the information from the future sequences via the "*forward layers*". The BiLSTM predicts the final state of action based on the forwarding and backward information, which is shown in Eqs. (13) and (14).

$$i_t^* = \sigma(w_i^* z_i^l + U_i^* h_{t+1} + b_i^*) \tag{13}$$

$$o_t^* = \sigma(v.i_t + v^*.i_t^*) \tag{14}$$

Next, we add the attention mechanism to recognize the most relevant parts of an action in the videos and produce a salient discriminative representation of the HAR. In the decoding phase, our attention mechanism selects the relevant part from the hidden vectors by utilizing the attention weights, which is represented in Eq. (16).

$$c_i = \sum_{j=1}^{T} \alpha_{ij} h_j \tag{15}$$

The hidden information between both the forward layer and the backward layer is split over time and generates useful output from the sequence by utilizing Eq. (17).

$$a_{ij} = \frac{\exp(w \cdot h_t)}{\sum_{k=1}^{t} \exp(w \cdot h_t)} \tag{16}$$

Here α in Eq. (15) shows the output probability of the current sequence. It represents the final state of action in a video sequence. The probability values are forwarded to a fully connected layer (FCN) for a high-level representation of the final output. For the final classification, we utilize both the Softmax loss and the center loss to enhance the model prediction, thereby improving its recognition accuracy.

3.4. Center loss

We compute the center loss together with the Softmax to produce the final probability and classify the actions collectively. The loss functions are calculated after the FCN layer. A fused loss method is applied to improve the prediction accuracy. With only the Softmax loss function, the model performance is lower due to a large distance within a class. We therefore utilize the mean λ setting to adjust the inter-class and intra-class distance to update the loss function. The center loss function simultaneously computes the loss among the features and their corresponding class centers [51]. The center loss function computes the minimum distance within-the class, whereas the Softmax loss calculates the maximum distance among the classes, as shown in Eqs. (17) and (18).

$$L_s = -\sum_{i=1}^{m} \log \frac{e^{w_{\mathrm{yi}}^T . x_i + b_{yi}}}{\sum_{j=1}^{n} e^{w_{\mathrm{yi}}^T . x_i + b_{yi}}} \tag{17}$$

$$L_c = \frac{1}{2} \sum_{i=1}^{m} \|x_i - c_{yi}\|_2^2 \tag{18}$$

In the above equations, "n" represents the classes, and "m" shows the batch size for the classifier. c_{yi} denotes the i^{th} sample of the class yi that belongs to the center of the class. We utilize the λ constraint for the center loss for updating c_{yi} and to avoid the misclassification in the real-time scenarios, as shown in Eq. (19). It balances both kinds of loss functions, achieving a better performance in action classification.

$$L = L_S + \lambda L_C \tag{19}$$

We compute both losses, $L_S + \lambda L_C$, find the fused information in L as a final loss function, and λ is used as a hyper parameter to adjust and update the loss as well as avoid the misclassification. We prove the superiority of the center loss function in the experimentation.

4. Experimental evaluation and discussion

In this section, we assessed the usefulness of the proposed system on three standard datasets: UCF11 [52], UCF Sports [53], and J-HMDB [54], showing better results than the SOTA methods. In the following subsections, we first described the datasets and the experimental settings in detail, and we then presented the achieved results. Finally, we demonstrated the comparison of the implemented system with the SOTA methods.

4.1. Datasets

UCF11 is a challenging dataset for video-based action recognition due to variation in illumination, cluttered background, and camera motions. The UCF11 dataset has a total of 1600 videos with eleven categories of action, such as shooting, jumping, riding, swimming etc., and all the videos are recorded at 30 frames per second (fps) rates.

The UCF Sports dataset has 150 sequences with 720 × 480 resolution videos that are collected from different sports like diving, riding horses, and golf swing, skateboarding, and lifting. The set of sports-related action videos is composed from various sources like BBC and ESPN, which are typically broadcast over television channels. These videos represent natural and real actions from different perspectives in a wide variety of scenes.

The J-HMDB dataset has 21 categories of different actions, such as catch, clamp, clap, hair, brush, baseball, swing, gunshot, jump, etc. It is a big video action collection that has 923 videos

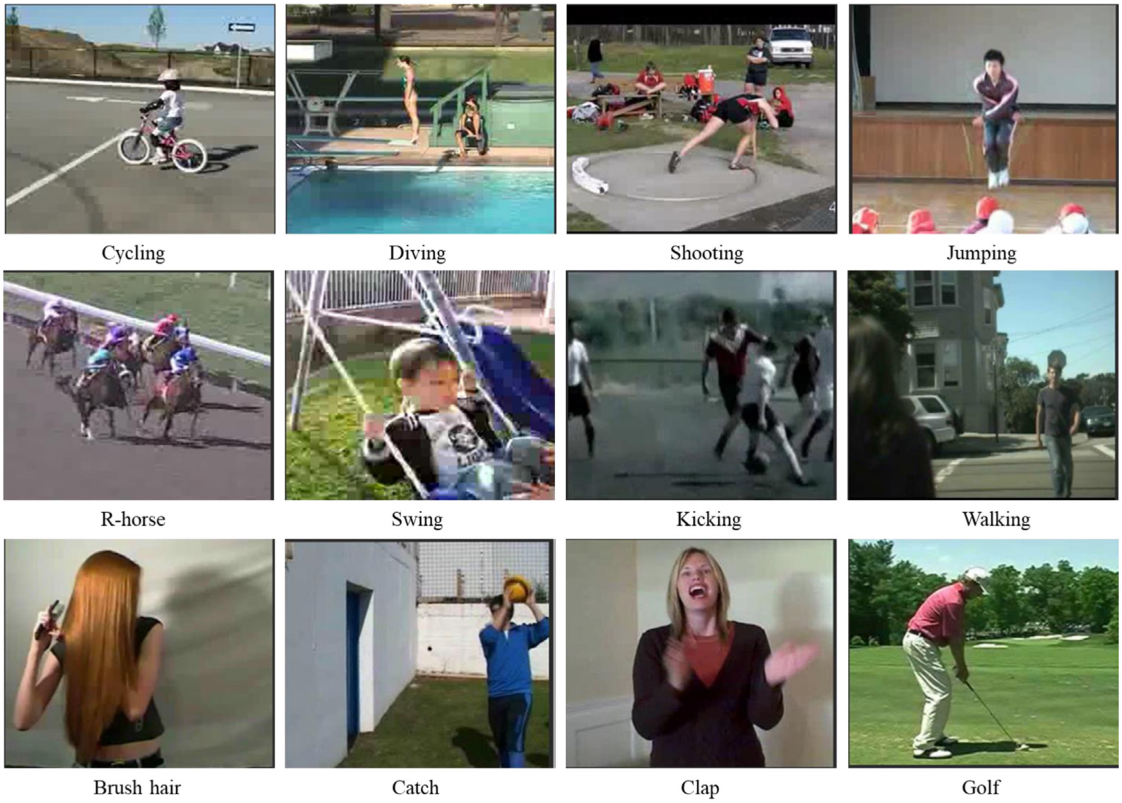

Fig. 3. Visual representations of sample action categories of UCF11, UCF Sports, and J-HMDB dataset.

with various actions, making it more challenging for the recognition task. Due to these challenges and problems, the J-HMDB achieved a poor performance compared to other datasets, but the recognition rate is better than the SOTA methods. The visual representation of sample actions from each dataset is given in Fig. 3.

4.2. Experimental results and discussion

We conducted extensive experimentation to evaluate and verify the efficiency of the HAR system for identifying actions in videos. We first validated the proposed system having both temporal and spatial features and compared it with a system that uses only sequential features. We further evaluated and compared our system having a spatio-temporal attention network with a spatial attention net where the attention network is applied after the convolutional process. We achieved better results with the suggested attention-based system than the other baseline attention methods, which implement either spatial or spatial–temporal information. It indicates the importance of temporal information in sequential data that can enhance the recognition performance, such as video-based action recognition. We conducted an ablation study for the best model configuration and demonstrated the importance and the significance of the DCNN for action recognition. The model architectures and the experimental evaluations are illustrated in Table 2.

Finally, we used the fusion strategy to fuse both types of features and scores for increasing the recognition rate in the classification. It is evident from Table 2 that the spatial and spatio-temporal attention features are more salient and discriminant than the spatial or the CNN features alone.

In the proposed system, we changed the traditional loss strategy and proposed a new fusion strategy by combining the center loss and the Softmax loss function to produce the best probabilities among the various actions. We also checked and assessed the efficacy of the center loss function with different values of lambda (λ), as shown in Fig. 6. We conducted different experiments to find the optimal λ value for increasing the recognition accuracy. The results obtained with various λ values are compared with the baseline approaches as presented in Table 3. The output of our system is superior to the other deep learning techniques on these datasets. Our system achieved a higher accuracy with the sports video, which is primarily due to these videos containing many similar activities that are difficult to recognize using a simple system. Our system learns deep spatial as well as temporal information to support its judgment in correctly identifying the actions within the sports videos. The class-level recognition results of the UCF11 and UCF sports datasets are illustrated in Fig. 4. Fig. 4 shows that our system recognized the basketball shooting, cycling, swimming, spiking, and juggling actions well with more than 98% accuracy. However, the system is confused among "*diving and jumping*" and "*walking and shooting*" which represent a poorer performance recognizing these actions due to their limited contextual differences.

The proposed system shows better results for the UCF sports dataset, where we achieved more than a 98% accuracy for all the categories except for the swing class, which is because of the system being mixed up between the *swing* class and the *swing bench* class due to their less contextual alteration. The results for the UCF sports and the UCF11 dataset are given in Fig. 4. The overall substitution of the recommended method is well aimed at the HAR task throught the experimentation.

Similarly, we evaluated our system on the J-HMDB challenging dataset for a better generalization of the proposed system to recognize tough actions like clapping, golf, jumping, standing, and walking. The suggested system recognized catch, sit, basketball, and run action with a more than 90% accuracy. The model achieved a high level of 80.2% overall classification on the J-HMDB challenging actions, proving the significance and the effectiveness

Table 2
The detailed ablation study of the proposed HAR system using the benchmark UCF11, UCF Sport, and J-HMDB datasets. The RB and SC represent the residual blocks and the skip connection in the model attributes.

Input	Proposed model architecture	UCF11 (%)	UCF Sports (%)	J-HMDB (%)
Video or Sequence of frames	Traditional CNN+BiLSTM+RB	82.20	83.80	75.81
	Traditional CNN+BiLSTM+RB+Attention	85.18	87.70	77.70
	Traditional CNN+BiLSTM+RB+Attention+Center Loss	85.93	89.59	77.90
	Dilated CNN+BiLSTM+RB	89.01	92.63	78.63
	Dilated CNN+BiLSTM+RB+Attention	96.31	97.24	79.24
	Dilated CNN+BiLSTM+RB+Attention+Center Loss	**98.30**	**99.10**	**80.20**

Table 3
Comparison of the proposed system with the SOTA models using the benchmark UCF Sports, UCF11, and J-HMDB datasets. The scores in bold font show the highest performances.

UCF sports		UCF11		J-HMDB	
Methods	Accuracy (%)	Methods	Accuracy (%)	Methods	Accuracy (%)
Tu et al. [34]	97.53	Patel et al. [37]	89.43	Tu et al. [34]	71.17
Meng et al. [16]	93.2	Meng et al. [16]	89.70	Ma et al. [46]	76.90
Gharaee et al. [44]	97.80	Gharaee et al. [44]	89.50	Ramsinghe et al. [55]	67.24
Gammulle et al. [35]	92.20	Gammulle et al. [35]	89.20	Gammulle et al. [35]	52.70
Nazir et al. [56]	97.30	Pan and Chao [57]	89.24	Ijjina et al. [36]	69.00
Dai et al. [14]	98.60	Dai et al. [14]	96.90	Dai et al. [14]	76.30
Our method (λ = 0.0)	98.50	Our method (λ = 0.0)	96.60	Our method (λ = 0.0)	**78.50**
Our method (λ = 0.01)	**99.10**	Our method (λ = 0.01)	**98.30**	Our method (λ = 0.01)	**80.20**

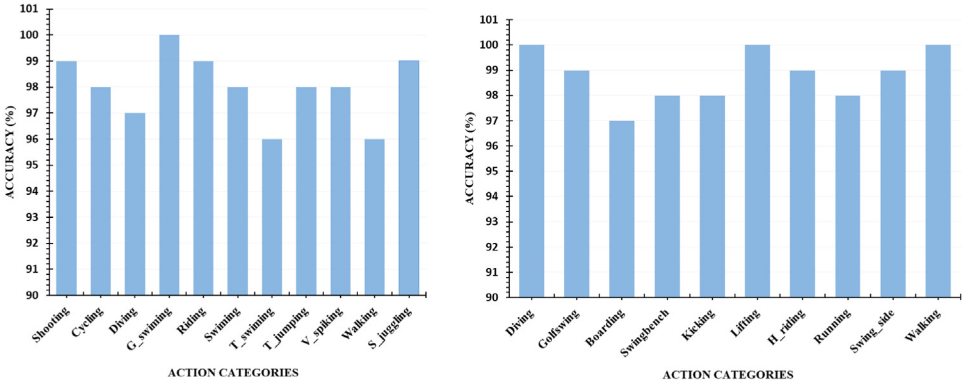

Fig. 4. The class-wise testing precision using the proposed system on the UCF11 dataset and the UCF Sports dataset.

of the proposed system. The class-wise recognition accuracies of the J-HMDB dataset are shown in Fig. 5.

Furthermore, we indicated the difference between the actual and the predicted label in the confusion matrix of the dataset. The confusion matrix of the UCF11 dataset is presented in Fig. 6(b). The confusion matrix represents the class level accuracy. The "*diving*" class achieved a high accuracy of 99.8%, and the "*walking*" class showed the lowest accuracy of 95.5% in the UCF11 dataset, respectively. Some action classes have similar motions that reduce the mutual error in the classification, such as volleyball spiking, basketball, and shooting.

The accuracy distribution in the confusion matrix of UCF Sports is moderately unchanged due to the precision rate among all the categories being higher than 95%, but the precision rate of the "*swing*" class is lower. We further investigated the reason for the worst performance of the system being in the "*swing*" class. After extensive experimentations, we found the main factor to be the swing side class and the swing bench class having less contextual information between each other, resulting in a poor classification. Our proposed system showed the discriminative ability for different action classes, which is evident from the J-HMDB confusion matrix illustrated in Fig. 7(b). We found that the actions like sit, run, swing, and catch achieved higher accuracies than clap and golf. Despite this, our proposed system can classify these actions with a better accuracy even for similar actions.

We deeply investigated our proposed action recognition system to obtain the prediction performance among the actual and the predicted classes, which are given in Fig. 7. The *x*-axis shows the predicted labels, and the *y*-axis shows the actual labels of the UCF Sport and the J-HMBD datasets in Fig. 7(a) and Fig. 7 (b). The actual recall values of all the actions are shown diagonally in the confusion matrices, and the confused classes are presented in the consistent rows of each class. For an efficient HAR system, we selected every eight frames from the video sequence and then extracted the highly discriminative features to ensure the real-time performance of the proposed model. We conducted various experimentations over different frames of the videos and reported them in Table 4. A higher prediction performance is reported on every eight frames, where the system skips the middle frames and jumps to every eight frame. The other jumps show worse performances over the proposed HAR system. The experimental outcomes of the proposed system illustrated a high generalization of the model for the action recognition, which outperformed other SOTA techniques (see Table 3).

5. Conclusion and future direction

Spatiotemporal features play an essential role in recognizing various actions in surveillance video data such as human action recognition. In this article, we proposed a unique attention-based pipeline for human action recognition, utilizing both the spatial and the temporal features from a sequence of frames. For this purpose, we used a CNN network to extract the high-level salient features from the video frames, and we then used the

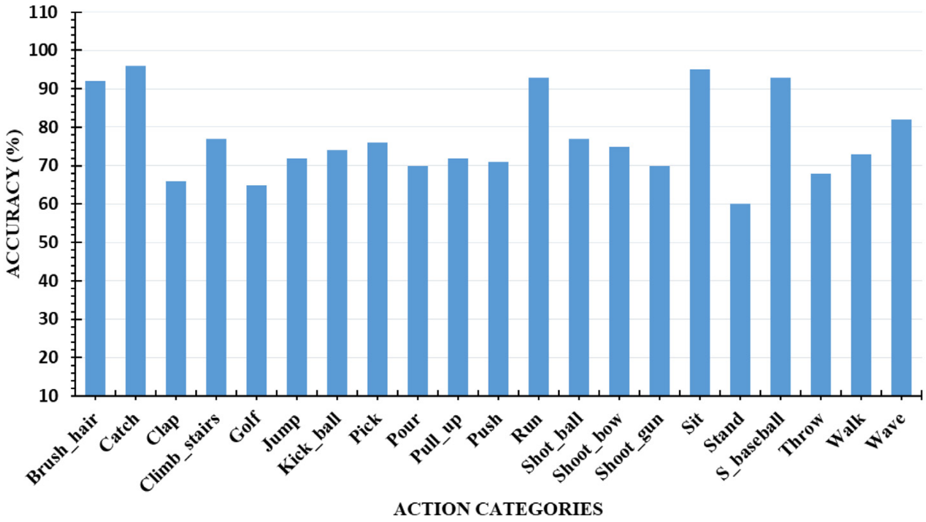

Fig. 5. The class-wise accuracy of the J-HMDB dataset using the proposed system.

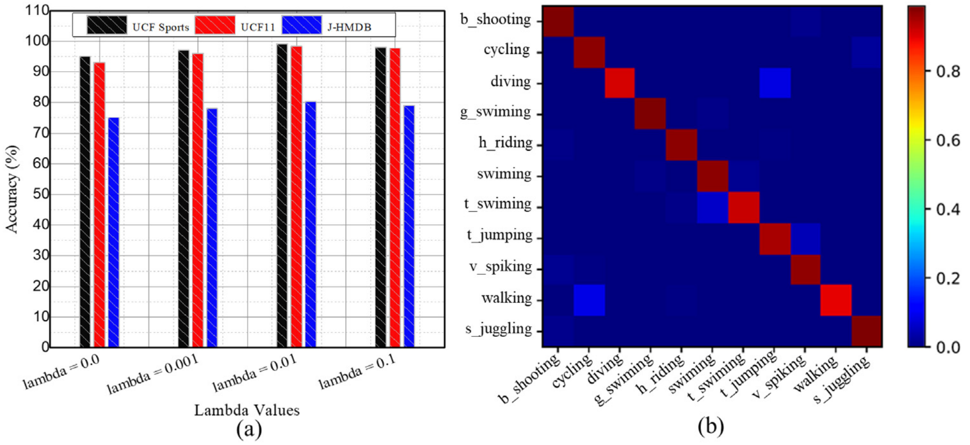

Fig. 6. Performance of the proposed system (a) with different values of lambda λ and (b) the confusion matrix of the UCF11 dataset among the actual and the predicted labels.

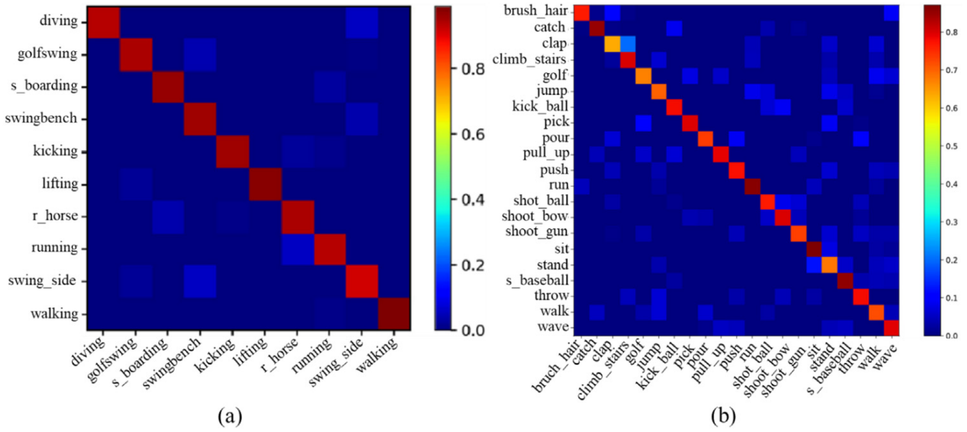

Fig. 7. The confusion matrices achieved using the proposed system on (a) the UCF Sports and (b) the J-HMDB datasets among the actual and the predicted labels.

skip connection approach to upgrade the learned features using the UFLBs and a dilated CNN. Furthermore, these spatial features were fed into the BiLSTM network to learn the temporal information. An attention layer is embedded to further determine the spatiotemporal information in more detail, which enhanced the performance at each step of the LSTM. The center and the softmax loss functions are employed to improve the classification performance of the human actions in the videos. We conducted extensive experiments on three standard benchmark datasets including the UCF11, the UCF Sports, and the J-HMDB. The proposed system achieved a recognition accuracy of 98.3% on the UCF11, 99.1% on the UCF Sports, and 80.5% on the J-HMDB dataset, reporting nearly 1%–3% improvements than the SOTA methods.

Table 4
Detailed experimental overview of frame selection using the proposed action recognition system, utilizing benchmark datasets including the UCF Sport, the UCF11, and the J-HMBD corpora. The scores in bold font show the highest performances.

Experiments	Frame selection	UCF sports (%)	UCF11 (%)	J-HMBD (%)
Test # 01	Jump to 04	93.15	89.00	75.00
Test # 02	Jump to 06	97.28	94.04	77.05
Test # 03	Jump to 08	**98.50**	**96.60**	**78.50**
Test # 04	Jump to 10	97.35	95.07	78.01

The proposed action recognition framework utilized a single stream learning strategy for human action recognition and it was evaluated over a medium scale HAR and interaction recognition datasets. In the future, we will extend our work by using a two-stream learning strategy to make it more intelligent for learning more discriminative features from the video frames to recognize complex actions in large-scale datasets. Our current system is flexible, and it can be helpful for adaptation in other domains, such as emotion and activity recognition, video summarization, and big data analytics.

CRediT authorship contribution statement

Khan Muhammad: Writing-original draft, Writing-review & editing, Supervision. **Mustaqeem:** Writing-original draft, Methodology, Writing-review & editing, Implementation. **Amin Ullah:** Investigation, Visualization. **Ali Shariq Imran:** Formal analysis, Project administration. **Muhammad Sajjad:** Resources, Funding acquisition, Supervision. **Mustafa Servet Kiran:** Writing - review & editing, Formal analysis. **Giovanna Sannino:** Revision, Conceptualization, Formal analysis. **Victor Hugo C. de Albuquerque:** Validation, Writing - review & editing.

Declaration of competing interest

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

Acknowledgments

This work was carried out during the tenure of an ERCIM ‘Alain Benoussan’ Fellowship Programme under the Contract 2019–40, Color and Visual Computing Lab at the Department of Computer Science, NTNU, Gjøvik, Norway. The work of Victor Hugo C. de Albuquerque was supported in part by the Brazilian National Council for Research and Development (CNPq) under Grant 304315/2017-6 and Grant 430274/2018-1.

All authors have read and agreed to the submitted version of the manuscript.

References

[1] N. Spolaôr, et al., A systematic review on content-based video retrieval, Eng. Appl. Artif. Intell. 90 (2020) 103557.
[2] A. Keshavarzian, S. Sharifian, S. Seyedin, Modified deep residual network architecture deployed on serverless framework of IoT platform based on human activity recognition application, Future Gener. Comput. Syst. 101 (2019) 14–28.
[3] A.D. Antar, M. Ahmed, M.A.R. Ahad, Challenges in sensor-based human activity recognition and a comparative analysis of benchmark datasets: A review, in: 2019 Joint 8th International Conference on Informatics, Electronics & Vision (ICIEV) and 2019 3rd International Conference on Imaging, Vision & Pattern Recognition, IcIVPR, IEEE, 2019.
[4] K.A. da Costa, et al., Internet of things: A survey on machine learning-based intrusion detection approaches, Comput. Netw. 151 (2019) 147–157.
[5] J.K. Aggarwal, M.S. Ryoo, Human activity analysis: A review, ACM Comput. Surv. 43 (3) (2011) 1–43.
[6] S. Pirbhulal, et al., Mobility enabled security for optimizing IoT based intelligent applications, IEEE Netw. 34 (2) (2020) 72–77.
[7] B. Ali, et al., A volunteer supported fog computing environment for delay-sensitive IoT applications, IEEE Internet Things J. (2020).
[8] S. Zhao, et al., Pooling the convolutional layers in deep convnets for video action recognition, IEEE Trans. Circuits Syst. Video Technol. 28 (8) (2017) 1839–1849.
[9] R. Girdhar, et al., Actionvlad: Learning spatio-temporal aggregation for action classification. in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017.
[10] R. Hou, C. Chen, M. Shah, An end-to-end 3d convolutional neural network for action detection and segmentation in videos, 2017, arXiv preprint arXiv:1712.01111.
[11] Y. Li, et al., Spatiotemporal interest point detector exploiting appearance and motion-variation information, J. Electron. Imaging 28 (3) (2019) 033002.
[12] C. Dai, et al., Human behavior deep recognition architecture for smart city applications in the 5G environment, IEEE Netw. 33 (5) (2019) 206–211.
[13] R. Khemchandani, S. Sharma, Robust least squares twin support vector machine for human activity recognition, Appl. Soft Comput. 47 (2016) 33–46.
[14] C. Dai, X. Liu, J. Lai, Human action recognition using two-stream attention based LSTM networks, Appl. Soft Comput. 86 (2020) 105820.
[15] H. Kwon, et al., First person action recognition via two-stream convnet with long-term fusion pooling, Pattern Recognit. Lett. 112 (2018) 161–167.
[16] B. Meng, X. Liu, X. Wang, Human action recognition based on quaternion spatial–temporal convolutional neural network and LSTM in RGB videos, Multimedia Tools Appl. 77 (20) (2018) 26901–26918.
[17] M. Baccouche, et al., Sequential deep learning for human action recognition, in: International Workshop on Human Behavior Understanding, Springer, 2011.
[18] D. Wu, N. Sharma, M. Blumenstein, Recent advances in video-based human action recognition using deep learning: a review, in: 2017 International Joint Conference on Neural Networks, IJCNN, IEEE, 2017.
[19] M. Alazab, et al., Intelligent mobile malware detection using permission requests and api calls, Future Gener. Comput. Syst. 107 (2020) 509–521.
[20] Y.-L. Hsueh, W.-N. Lie, G.-Y. Guo, Human behavior recognition from multiview videos, Inform. Sci. (2020).
[21] M. Elhoseny, et al., A hybrid model of internet of things and cloud computing to manage big data in health services applications, Future Gener. Comput. Syst. 86 (2018) 1383–1394.
[22] X. Zhen, L. Shao, Action recognition via spatio-temporal local features: A comprehensive study, Image Vis. Comput. 50 (2016) 1–13.
[23] B. Saghafi, D. Rajan, Human action recognition using pose-based discriminant embedding, Signal Process., Image Commun. 27 (1) (2012) 96–111.
[24] J. Lee, H. Jung, TUHAD: Taekwondo unit technique human action dataset with key frame-based CNN action recognition, Sensors 20 (17) (2020) 4871.
[25] H. Yasin, M. Hussain, A. Weber, Keys for action: An efficient keyframe-based approach for 3D action recognition using a deep neural network, Sensors 20 (8) (2020) 2226.
[26] Y. Zhao, et al., Multi-feature fusion action recognition based on key frames, in: 2019 Seventh International Conference on Advanced Cloud and Big Data (CBD), IEEE, 2019.
[27] X. Wei, et al., Piecewise classifier mappings: Learning fine-grained learners for novel categories with few examples, IEEE Trans. Image Process. 28 (12) (2019) 6116–6125.
[28] A. Garcia-Garcia, et al., A survey on deep learning techniques for image and video semantic segmentation, Appl. Soft Comput. 70 (2018) 41–65.
[29] J. Schmidhuber, Deep learning in neural networks: An overview, Neural Netw. 61 (2015) 85–117.
[30] T.M. Lee, J.-C. Yoon, I.-K. Lee, Motion sickness prediction in stereoscopic videos using 3D convolutional neural networks, IEEE Trans. Vis. Comput. Graphics 25 (5) (2019) 1919–1927.
[31] S.U. Khan, et al., Cover the violence: A novel deep-learning-based approach towards violence-detection in movies, Appl. Sci. 9 (22) (2019) 4963.
[32] K. Simonyan, A. Zisserman, Two-stream convolutional networks for action recognition in videos, in: Advances in Neural Information Processing Systems, 2014.
[33] C. Feichtenhofer, A. Pinz, A. Zisserman, Convolutional two-stream network fusion for video action recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016.
[34] Z. Tu, et al., Multi-stream CNN: Learning representations based on human-related regions for action recognition, Pattern Recognit. 79 (2018) 32–43.
[35] H. Gammulle, et al., Two stream lstm: A deep fusion framework for human action recognition, in: 2017 IEEE Winter Conference on Applications of Computer Vision, WACV, IEEE, 2017.
[36] E.P. Ijjina, C.K. Mohan, Hybrid deep neural network model for human action recognition, Appl. Soft Comput. 46 (2016) 936–952.

[37] C.I. Patel, et al., Human action recognition using fusion of features for unconstrained video sequences, Comput. Electr. Eng. 70 (2018) 284–301.

[38] R.R. Guimaraes, et al., Intelligent network security monitoring based on optimum-path forest clustering, Ieee Netw. 33 (2) (2018) 126–131.

[39] C. Xu, et al., Redundancy avoidance for big data in data centers: A conventional neural network approach, IEEE Trans. Netw. Sci. Eng. 7 (1) (2018) 104–114.

[40] X. He, et al., Green resource allocation based on deep reinforcement learning in content-centric IoT, IEEE Trans. Emerg. Top. Comput. 8 (3) (2018) 781–796.

[41] S. Kulkarni, S. Jadhav, D. Adhikari, A survey on human group activity recognition by analysing person action from video sequences using machine learning techniques, in: Optimization in Machine Learning and Applications, Springer, 2020, pp. 141–153.

[42] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, 2014, arXiv preprint arXiv:1409.0473.

[43] J. Wen, et al., Big data driven marine environment information forecasting: A time series prediction network, IEEE Trans. Fuzzy Syst. (2020).

[44] Z. Gharaee, P. Gärdenfors, M. Johnsson, First and second order dynamics in a hierarchical SOM system for action recognition, Appl. Soft Comput. 59 (2017) 574–585.

[45] J. Chen, Z. Lv, H. Song, Design of personnel big data management system based on blockchain, Future Gener. Comput. Syst. 101 (2019) 1122–1129.

[46] M. Ma, et al., Region-sequence based six-stream CNN features for general and fine-grained human action recognition in videos, Pattern Recognit. 76 (2018) 506–521.

[47] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780.

[48] J.G. Zilly, et al., Recurrent highway networks, in: Proceedings of the 34th International Conference on Machine Learning, Vol. 70, 2017, JMLR. org.

[49] M. Arsalan, et al., OR-Skip-net: Outer residual skip network for skin segmentation in non-ideal situations, Expert Syst. Appl. 141 (2020) 112922.

[50] N. Khan, et al., SD-Net: Understanding overcrowded scenes in real-time via an efficient dilated convolutional neural network, J. Real-Time Image Process. (2020) 1–15.

[51] W. Xiong, et al., A discriminative feature learning approach for remote sensing image retrieval, Remote Sens. 11 (3) (2019) 281.

[52] J. Liu, J. Luo, M. Shah, Recognizing realistic actions from videos in the wild, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2009.

[53] L. Shao, et al., Spatio-temporal Laplacian pyramid coding for action recognition, IEEE Trans. Cybern. 44 (6) (2013) 817–827.

[54] H. Jhuang, et al., Towards understanding action recognition, in: Proceedings of the IEEE international conference on computer vision, 2013.

[55] S. Ramasinghe, et al., Combined static and motion features for deep-networks based activity recognition in videos, IEEE Trans. Circuits Syst. Video Technol. (2017).

[56] S. Nazir, et al., A bag of expression framework for improved human action recognition, Pattern Recognit. Lett. 103 (2018) 39–45.

[57] Z. Pan, C. Li, Robust basketball sports recognition by leveraging motion block estimation, Signal Process., Image Commun. (2020) 115784.

Khan Muhammad (S'16–M'18) is an Assistant Professor at the Department of Interaction Science and Director of Visual Analytics for Knowledge (VIS2KNOW) Laboratroy, Sungkyunkwan University, Seoul, South Korea. His research interests include intelligent video surveillance (fire/smoke scene analysis, safe transportation systems, and disaster management), medical image analysis (brain MRI, diagnostic hysteroscopy, and wireless capsule endoscopy), information security (steganography, encryption, watermarking, and image hashing), video summarization (single-view and multi-view), multimedia, computer vision, IoT, smart cities, and video analytics. He has filed 8 patents and has published over 170 papers in peer-reviewed international journals and conferences in these research areas.

Mustaqeem received the bachelor's degree in computer science from Institute of Business and Management Sciences (IBMS), Agriculture University Peshawar, Pakistan and took his Master's degree from Islamia College, Peshawar, Pakistan with research in video analysis (Content base video retrieval/Action recognition). He is currently pursuing Ph.D. degree with Interaction Technology Laboratory, in digitals contents from College of Electronics and Information Engineering, Sejong University, Seoul, Republic of Korea. He is working as a researcher at Interaction Technology Laboratory (IT Lab). His major research interests include Audio digital signals processing, speech processing, Emotion recognition, image processing, and video processing.

Amin Ullah received Ph.D. degree in digital contents from Sejong University, South Korea. He is currently working as a Postdoc Researcher at the CoRIS Institute, Oregon State University, Corvallis 97331, Oregon, USA. His major research focus is on human action and activity recognition, sequence learning, image and video analytics, content-based indexing and retrieval, IoT and smart cities, and deep learning for multimedia understanding. He has published several papers in reputed peer reviewed international journals and conferences including IEEE Transactions on Industrial Electronics, IEEE Transactions on Industrial Informatics, IEEE Transactions on Intelligent Transportation Systems, IEEE Internet of Things Journal, IEEE Access, Elsevier Future Generation Computer Systems, Elsevier Applied Soft Computing, International Journal of Intelligent Systems, Springer Multimedia Tools and Applications, Springer Mobile Networks and Applications, and IEEE Joint Conference on Neural Networks.

Ali Shariq Imran received the master's degree in software engineering and computing from the National University of Science & Technology (NUST), Pakistan, in 2008, and the Ph.D. degree in computer science from the University of Oslo (UiO), Norway, in 2013. He is currently an Associate Professor with the Department of Computer Science, Norwegian University of Science and Technology (NTNU), Norway. He specializes in applied research focusing on deep learning technology and its application to signal processing, natural language processing, and the semantic Web. He has over 65 peer-reviewed journals and conference publications to his name. He has served as a reviewer for many reputed journals over the years. He is a member of the Norwegian Colour and Visual Computing Laboratory (Colourlab) and the IEEE Norway Section.

Muhammad Sajjad received the master's degree from the Department of Computer Science, College of Signals, National University of Sciences and Technology, Rawalpindi, Pakistan in 2012, and the Ph.D. degree in digital contents from Sejong University, Seoul, South Korea in 2015. He is currently working as an ERCIM Research Fellow at NTNU, Norway. He is an Associate Professor with the Department of Computer Science, Islamia College University Peshawar, Pakistan. He is also the Head of the Digital Image Processing Laboratory with Islamia College University Peshawar, where many students are involved in different research projects under his supervision, such as Big data analytics, medical image analysis, multi-modal data mining and summarization, image/video prioritization and ranking, fog computing, the Internet of Things, autonomous navigation, and video analytics. His primary research interests include computer vision, image understanding, pattern recognition, robotic vision, and multimedia applications, with current emphasis on economical hardware and deep learning, video scene understanding, activity analysis, fog computing, the Internet of Things, and real-time tracking. He has published more than 65 papers in peer-reviewed international journals and conferences. He is serving as a professional reviewer for various well-reputed journals and conferences. Currently, he is the associate editor at IEEE Access and acting as a guest editor at IEEE Transactions on Intelligent Transportation Systems.

Mustafa Servet Kiran received the B.S. and Ph.D. degrees in computer engineering from the Institute of Natural and Applied Sciences, Selcuk University, Konya, Turkey, in 2010 and 2014, respectively. He is currently an Associate Professor with the Computer Engineering Department, Konya Technical University. His current research interests include swarm intelligence, evolutionary algorithms, and their real-world applications.

Giovanna Sannino (M'09) received the BachBachelor's degree in computer engineering from the University of Naples Federico II, Naples, in 2008, and the Master's degree, named "European Master on Critical Networked Systems" from the University of Naples Parthenope, Naples, Italy, in 2009. Successively, she received the Master's degree in telecommunications engineering (Cum Laude) and the Ph.D. degree in information engineering from the University of Naples Parthenope, Naples, Italy, in April 2011 and April 2015, respectively. She is currently a Researcher with the Institute of High-Performance Computing and Networking, of the National Research Council of Italy, Naples. Her research interests include the areas of mobile Health, pervasive computing, pattern recognition, signal processing, and artificial intelligence for healthcare. Dr. Sannino is the author of more than 70 papers in international journals and in the proceedings of international conferences, and she is serving as a professional reviewer for many well-reputed journals and prestigious conferences over the years. She has been involved as Steering Committee member, workshop Chair and Session Chair in the organization of several international/national conferences, and has been Guest Editor of several special issues on international journals. Dr. Sannino is a Member of the IEEE 11073 Personal Health Device Working Group and Young Researchers Committee Member of the World Federation on Soft Computing, and is an Associate Editor for the Biomedical Signal Processing and Control (Elsevier).

Victor Hugo C. de Albuquerque [M'17, SM'19] is a collaborator Professor and senior researcher at the Graduate Program on Teleinformatics Engineering at the Federal University of Ceará, Brazil, and at the Graduate Program on Electrical Engineering, Federal University of Ceará, Fortaleza/CE, Brazil. He has a Ph.D in Mechanical Engineering from the Federal University of Paraíba (UFPB, 2010), an MSc in Teleinformatics Engineering from the Federal University of Ceará (UFC, 2007), and he graduated in Mechatronics Engineering at the Federal Center of Technological Education of Ceará (CEFETCE, 2006). He is a specialist, mainly, in Image Data Science, IoT, Machine/Deep Learning, Pattern Recognition, Robotic.