Hindawi
Computational Intelligence and Neuroscience
Volume 2022, Article ID 3454167, 10 pages
https://doi.org/10.1155/2022/3454167

Research Article

Vision Transformer and Deep Sequence Learning for Human Activity Recognition in Surveillance Videos

Altaf Hussain, Tanveer Hussain, Waseem Ullah, and Sung Wook Baik

Sejong University, Seoul 143-747, Republic of Korea

Correspondence should be addressed to Sung Wook Baik; sbaik3797p@gmail.com

Received 26 October 2021; Revised 9 December 2021; Accepted 18 December 2021; Published 4 April 2022

Academic Editor: Bai Yuan Ding

Human Activity Recognition is an active research area with several Convolutional Neural Network (CNN) based features extraction and classification methods employed for surveillance and other applications. However, accurate identification of HAR from a sequence of frames is a challenging task due to cluttered background, different viewpoints, low resolution, and partial occlusion. Current CNN-based techniques use large-scale computational classifiers along with convolutional operators having local receptive fields, limiting their performance to capture long-range temporal information. Therefore, in this work, we introduce a convolution-free approach for accurate HAR, which overcomes the above-mentioned problems and accurately encodes relative spatial information. In the proposed framework, the frame-level features are extracted via pretrained Vision Transformer; next, these features are passed to multilayer long short-term memory to capture the long-range dependencies of the actions in the surveillance videos. To validate the performance of the proposed framework, we carried out extensive experiments on UCF50 and HMDB51 benchmark HAR datasets and improved accuracy by 0.944% and 1.414%, respectively, when compared to state-of-the-art deep models.

1. Introduction

Nowadays, surveillance systems are increasingly installed for monitoring purposes to ensure public safety and put an effort to mitigate crimes [1, 2]. Therefore, enormous amounts of data are generated on the daily basis from the CCTV cameras, requiring manual monitoring for undesired activities [3]. It is almost impossible for human being to monitor multiple video streams, particularly, simultaneous inspection of several cameras for the identification of different activities and events. Therefore, automating the process of Human Activity Recognition (HAR) using image processing and deep learning techniques offers promising solution to this problem. Computer vision techniques for HAR in surveillance system are reliable sources for automatic decision-making, which are responsible for identifying individuals performing suspicious activities and informing the law enforcement agencies to take appropriate preemptive actions. Besides smart surveillance, HAR has numerous applications such as video retrieval [4] and video summarization [5]. However, accurate HAR using computer vision techniques is challenging due to instantaneous transition of events in successive frames, illumination variations, different viewpoints, cluttered background, and different scaling [6]. In videos context, activity recognition relies on the collection of multiple consecutive video frames, where both spatial and temporal information need to be analysed for an individual's body movements. The HAR related literature can be broadly divided into traditional features and deep learning-based techniques, as discussed in the subsequent sections. The traditional methods mainly consist of three steps: (1) preprocessing step is used to remove outliers and noise, (2) the preprocessed data undergo several low-level features extraction phases, and (3) in classification step, the extracted features are intelligently mapped with their corresponding classes. Since our model is based on deep features, therefore, herein, we only discuss deep features-based methods, but interested readers are referred to extensive reviews of baseline HAR methods [7, 8].

Deep Learning-Based Techniques. Considering the limited performance of traditional features and machine learning methods in HAR domain, researchers focus on deep learning-based techniques that process data in an end-to-end manner for features extraction and classification. CNN-based techniques extract features in a hierarchical way, where initial layers extract local features and the final layers extract global features. The standard 2D CNN is effective in learning the spatial information but is unable to learn the temporal information, which is important for improved accuracy in the case of HAR. For instance, Karpathy et al. [9] used pretrained CNN model which fuses the information from multiple frames for efficient activity recognition. This work is extended by Simmoyan and Zisserman [10], where the authors proposed a two-stream CNN network to overcome the issue of motion information involvement in HAR by employing optical flow features. But, in this method, the long-range contextual information is not considered, as the final activity prediction is generated by taking the average of predictions from a segmented video shot of variable length, that is, 10 to 15 frames sequence. Recently, simple 2D CNN architecture is widely employed in HAR literature; for instance, Khan et al. [11] used a lightweight CNN MobileNet model to cover violent activity scenes in the movies. Similarly, in another research, the Hough forests features are combined with 2D CNN to train a hybrid model for violent activity recognition [12]. The authors claim that their method requires less computation time, but due to complex poses, different scales, and illuminations, such methods are not effective in real-world situations. In this regard, the 2D CNN-based methods and motion representation attributes are gradually upgraded from 2D to 3D CNN, which considers both spatial and temporal information [13]. Researchers thus introduced 3D CNN and its variants to capture spatio-temporal information. In this direction, Tran et al. [14] proposed a 3D CNN (C3D) for the temporal and spatial features representation in the video data, outperforming the existing methods for HAR. In another research, Carreira et al. [15] proposed a novel mechanism where the pretrained ImageNet 2D filters are modified into their corresponding 3D version for activity recognition. The experiments concluded that their method achieved higher accuracy as compared to randomly initialized filters. Similarly, Hussain et al. [16] have proposed a lightweight 3D CNN model for anomaly activity recognition and camera prioritization in surveillance environments. The variants of 3D CNNs include two-stream 3D CNN [17], pseudo-3D CNN [18], and MiCT-Net [19]. However, the existing 3D-CNN models can only process 10 to 16 frames effectively. They cannot recognize lengthy activities due to exponential increase in time complexity caused mainly by the temporal dimension [16]. To overcome this issue, researchers experimented on hybrid models where the spatial features are extracted from pretrained CNN models and learned the temporal information using variants of Recurrent Neural Networks (RNNs).

The hybrid models extract frame-level features using CNNs or optical flow models, which are stacked using a predefined time stamp to learn temporal information that are fed to RNN variants such as long short-term memory (LSTM) Network [20] and Gated Recurrent Unit (GRU) [21] to improve the HAR performance. For sequence learning, the hybrid models have used RNN, LSTM, and GRU but without focusing on the selective information in the spatial domain from the consecutive sequences, which is very important to maintain connectivity between the frames. However, extracting selective features and discriminative information for HAR in complex surveillance environments [22] is a challenging task. Therefore, researchers proposed different techniques; for example, in an existing research, Li et al. [23] extracted C3D features from the input video sequence by using sliding window techniques to generate cubes, which are fed to LSTM network for HAR. Ma et al. [24] utilized VGG19 spatial features in their framework and employed multilayer LSTM for optimal activity recognition. The AlexNet CNN model is used in a baseline research [25] for spatial features extraction, and then deep bidirectional LSTM is used for temporal learning. The authors in [26] have proposed an efficient approach for real-time HAR in surveillance systems by using MobileNet architecture for spatial features extraction followed by sequential learning with LSTM. More recently, Hussain et al. [27] have achieved significant performance by integrating CNN features with support vector machine as sequential learning model, which is not considered to be more robust and generalized towards patterns learning when compared to recent deep sequential models.

In the existing methods, the kernels of the CNNs are mainly designed to capture short-range spatial-temporal information and they are limited to learn the long-range dependencies when they are beyond their receptive field. However, stacking convolutional layers [10, 13], naturally, extends the receptive field, but these strategies are inherently limited in capturing long-range dependencies by means of aggregation of shorter-range information. Still, the long-range temporal dependency is unresolved because the above-mentioned approaches are strongly relying on the weak features selection [27]. Similarly, capturing long-range sequences dependency is a major problem in different domains such as natural language processing (NLP) in machine translation [28, 29], autoregressive word generation [30], and question answering [31]. Therefore, the field of NLP has been revolutionized with the emerging technique such as self-attention or transformer [32]. Activity recognition and NLP share several high-level similarities; for example, sentences and videos both are sequential forms of data, where a single word is insufficient to understand its contextual meaning in terms of a sentence similar to a video where a single frame is not enough to understand semantics of the whole video. In a baseline research for Vision Transformer (ViT) [33], the authors extracted the local spatial features from an input image using patching strategy and then encoded these features to a standard transformer from the NLP with novel modifications; as a result, they achieved superior performance against state-of-the-art for classification tasks. Therefore, in this work, we investigated that long-range self-attention model would be highly effective in HAR. The problem of learning long-range spatiotemporal features in the HAR is resolved by discriminative spatial features extraction via ViT. Therefore, the frame-level features are extracted from the pretrained ViT-Base-16 followed by LSTM for the HAR. The main contributions of the proposed method are as follows:

(1) We propose a novel mechanism that utilizes CNN-free approach to capture surveillance videos long-range temporal dependencies using ViT, followed by a sequential learning method to achieve new state-of-the-art accuracy when compared to existing HAR methods.

(2) Spatial and temporal features play an important role in the accurate HAR, where we employ ViT for spatial features and multilayered LSTM to learn temporal relationships among these features to recognize human activities with higher confidence.

(3) The performance of the proposed framework is evaluated on the challenging HMDB51 and UCF50 HAR datasets. The experimental results accomplished new state-of-the-art accuracy of 73.714% using HMDB51 and 96.144% accuracy using UCF50.

The rest of the paper is organized as follows: Section 2 presents the proposed activity recognition framework. Experimental setup, datasets, discussion on the results, and comparative analysis are given in Section 3. The conclusion and future works are given in Section 4.

2. The Proposed Activity Recognition Framework

The proposed framework for HAR mainly consists of three steps, as visualized in Figure 1. In the first step, surveillance cameras capture video streams which are then fed to pretrained ViT-Base-16 for frame-level spatial features extraction. The spatial features are stacked together to create a resultant feature vector from 30 consecutive frames. In the third step, the generated feature vector is fed into a multilayer LSTM network to capture long-range spatial-temporal dependency in the HAR.

2.1. Features Extraction Using Vision Transformer. The architecture of ViT-Base-16 is entirely based on the standard transformer [32] architecture and achieved remarkable accuracy when compared to CNN-based models for image classification tasks. It uses self-attention mechanism to capture long-range relationship between input sequences. ViT is actually an attempt to use transformer model for image classification. Basically, it divides the input image into a number of patches that are linearly projected with learnable positional embedding to learn the order of patches followed by transformer encoder with multilayer perceptron for final classification.

In the first part, the input image is divided into non-overlapping patches, because a standard transformer receives 1D sequence of token as an input. Usually, the image is in 2D format; therefore, to handle the 2D image, an image $x \in \mathbb{R}^{H\times W\times C}$ is reshaped into a sequence of flattened 2D patches $x_p \in \mathbb{R}^{N\times (P^2.C)}$. Herein, (H, W, C) represents the height, width, and channels of the image, while (P, P) is the resolution of each image patch, and $(N = HW/P^2)$ is the total number of patches. Typically, the patch size P is chosen as 16×16 *or* 32×32, where the small P size is able to capture longer sequences and vice versa. In our case, we have used the $16\times 16\,P$ for features extraction; in the subsequent section, these submodules are discussed in detail.

2.2. Linear Embedding Layer. The sequence patches are linearly projected into a vector with dimension d using a learn embedding matrix E. Then, these embedded representations are concatenated together with learnable classification token v_{class}. The embedded patches are without order; therefore, positional information E_{pos} is used to re-order the spatial information as the original image. The result of embedded patches with token Z_0 is mathematically represented in (1).

2.3. Vision Transformer Encoder. The resultant embedding patches of Z_0 (in (1)) are fed to the transformer encoder module, which consists of L identical layers as shown in Figure 2. Furthermore, each module is divided into two components such as multihead self-attention (MSA) block and multilayer perceptron (MLP). The last block of MLP consists of two dense layers. Equations (2) and (3) represent the mathematical representations of the MSA and MLP, respectively.

$$\begin{aligned} Z_0 = {} & \left[v_{calss}; x_1E; x_2E; \,..x_nE\right] \\ & + E_{pos}, E \in \mathbb{R}^{\left(P^2.C\right)\times d}, E \in \mathbb{R}^{(n+1)\times d}. \end{aligned} \tag{1}$$

$$z_l' = MSA\left(LN\left(z_{l-1}\right)\right) + z_{l-1}, \, l = 1\ldots L, \tag{2}$$

$$z_l = MLP\left(LN\left(z_l'\right)\right) + z_l', \, l = 1\ldots L. \tag{3}$$

In the last layer of the encoder, the first element z_L^0 in the sequence is passed to external head classifier for predicting the class label.

$$y = LN\left(z_L^0\right). \tag{4}$$

The MSA is the central component of the transformer model which calculates the most and least important patch and discard the later one from the input sequence. It is further divided into four layers such as linear, self-attention, and concatenation layers to combine the output from the multiple heads, as their graphical representation is shown in Figure 2(c). Basically, the attention mechanism can be represented by attention weights which is calculated from the weighted sum of all values in sequences z. Three values,

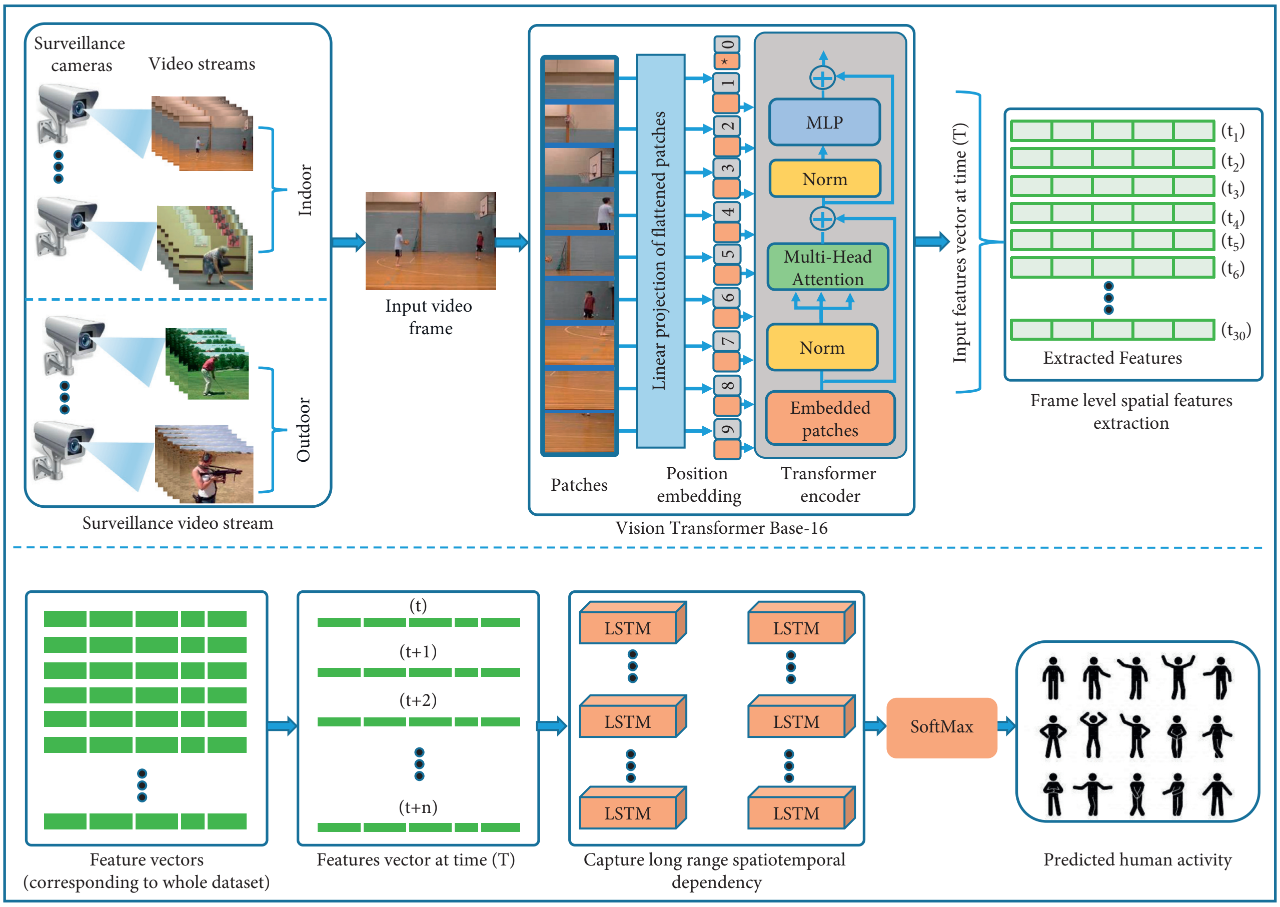


Figure 1: The proposed framework for automatic activity recognition in surveillance videos.

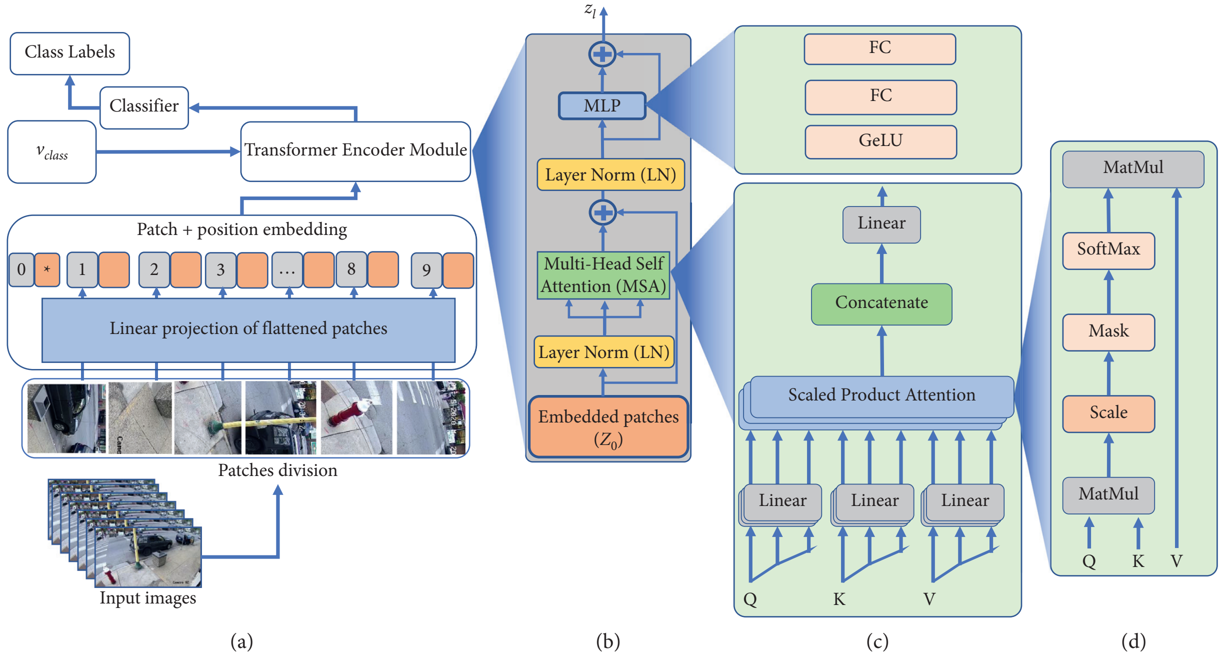


Figure 2: The Vision Transformer architecture: (a) the main architecture of the model, (b) the transformer encoder module, (c) multiscale self-attention (MSA) head, and (d) the self-attention (SA) head.

Q (query), K (key), and V (value), are calculated from the input sequence by multiplying the elements (Q, K) against three learning matrices U_{QKV}; a single SA is graphically represented in Figure 2(d), while the mathematical formulation is given in equation (5):

$$[Q, K, V] = zU_{QKV}, U_{QKV} \in \mathbb{R}^{d\times 3D_K}. \tag{5}$$

In a given input sequence, to calculate the importance of one element with respect to others, the value of Q vector is multiplied by dot product with the K vectors. Then their result is scaled and passed to SoftMax activation function to find out the importance of patch with high attention score, as given mathematically in equation (6):

$$A = SoftMax\left(\frac{QK^T}{\sqrt{D_K}}\right), \ A \in \mathbb{R}^{n\times n}. \tag{6}$$

Table 1: Parameters details shown in the formulation of LSTM network.

Variables/symbol	Description
Δt	Input over time t
∂	Sigmoid activation function
w	Weights
b	Bias terms
i_t	Input gate
f_t	Forget gate
	Output gate
tanh	Tan h activation function
SoftMax	Activation for the final classification
N	Numbers of classes

The MSA is actually the combination of the multiple attention heads h instead of single values of Q, K, and V. For robust and optimal features selection, the results of each SA are concatenated and then projected through a feedforward layer with learnable weights W to the desired dimensions, as expressed in equation (7):

$$MSA(z) = Concat\,(SA_1(z); SA_2(z); \ldots SA_h(z))W, W \in \mathbb{R}^{h.D_K\times D}. \tag{7}$$

2.4. Learning Long-Range Temporal Dependencies via LSTM.

Temporal features are very important to learn long-range dependencies in activity recognition. The RNN model is specifically designed for the time series or continuous data but recently researchers are inspired by their performance in activity recognition domain [34]. It combines the learned information from the previous and the current frames in an input video sequence for accurate HAR identification. However, the RNN is unable to hold long-range temporal dependencies due to vanishing gradient problem, which is solved by LSTM [35], that is capable of holding long-range temporal information. In the LSTM architecture, there are three gates, (1) input, (2) output, and (3) forget gates. Table 1 shows the parameters details used to formulate the internal mechanism of LSTM to capture long-range temporal dependency for HAR.

The last gate f_t is responsible for retaining or discarding irrelevant information in input Δt and from the previous output S_{t-1} [36, 37]. The frame-level discriminative features from the ViT at the unit time, t, are passed into the stacked LSTM network and f_t to hold long-term temporal dependency. Equations (8) to (14) show the mathematical representation of the LSTM network.

$$i_t = \partial(w_i[\Delta t + S_{t-1}] + b_i), \tag{8}$$

$$f_t = \partial(w_f[\Delta t + S_{t-1}] + b_f), \tag{9}$$

$$O_t = \partial(w_0[\Delta t + S_{t-1}] + b_0), \tag{10}$$

$$R = \tanh(w_R[\Delta t + S_{t-1}] + b_R), \tag{11}$$

$$C_t = C_{t-1}.f_t + R.i_t, \tag{12}$$

$$S_t = tahn(C_t).\, O_t, \tag{13}$$

$$\mathrm{P}rediction_{state} = soft\max(s_t N). \tag{14}$$

Herein, the term Δt represents the input over time and sigmoid activation function is represented by ∂. Their weights and bias terms are represented by w and b, respectively. The forget gate f_t at time t keeps the information of the previous frame that is needed and discard it otherwise. The output gate O_t keeps the information of the upcoming step and R is the recurrent unit having tanh activation function. It is computed from the input of the current frame and state of the previous frame S_{t-1}. The RNN hidden state is calculated by the tanh activation and memory cell C_t. The activity recognition does not need intermediate output from the LSTM; therefore, we use the SoftMax activation for final classification in (14), and N is used to represent the number of classes, that is, 51 for HMDB51 and 50 for UCF50 dataset.

2.5. Modeling Human Activity Recognition via ViT and Multilayer LSTM.

Recently, ViT have dominated CNNs for image representation, leading to better classification [33] and segmentation results [38]. Herein, inspired by the better representation abilities of ViT, we extract spatial transformer features in our framework using pretrained ViT model, followed by sequential learning method to learn the temporal dependencies and interpretations of input frames. There are different variants of ViT models such as ViT-Base, ViT-Large, and ViT-Huge. The ViT-Base-16 achieved remarkable accuracy when compared to existing image classification methods using benchmark datasets, indicating its robust and representative features potentials. The ViT models have varied number of encoder layers, hidden dimension size, number of attention heads used by MSA layer, and MLP classifier size, as detailed in Table 2, where we employ ViT-Base-16 model with 16×16 patch size in the proposed framework.

Due to complex patterns of actions and temporal gaps between sequential actions merged together to form single activity, a single LSTM cell is unable to learn the patterns accurately. Therefore, we have conducted multiple experiments and stacked multiple LSTM cells to learn long-term temporal patterns across the video sequences. Table 3 shows the proposed LSTM network to capture long-range temporal dependency of activity recognition. The ViT-Base-16 extracts 1000 features vector from each frame; therefore, our proposed sequential learning model takes 30 frames with 1000 spatial features vector.

Initially, the features vector contains enriched patterns information; therefore, we have used 128 LSTM units to learn all possible discriminative features. Then the features space is reduced by 64 numbers of LSTM units to efficiently map the number of classes, that is, 51 and 50 classes of HMDB51 and UCF50, respectively. Furthermore, to avoid overfitting and make the network more stable during training with faster learning, we utilized a 50% dropout and batch normalization. On top of this, we also performed experiments using different learning rates (LR) because it is one of the most important hyperparameters, which greatly affects the generalization of the model. When we take 1e-1 LR, the weight of the model is updated drastically and causes overshoot due to which the model does not reach global minima and is stuck in the local minima. After multiple experiments, when we take 1e-4 LR, the model achieves the highest accuracy of 73.714% and 96.144% on the HMDB51 and UCF50 datasets, respectively.

3. Experimental Results and Discussion

The performance of the proposed framework is validated over two benchmark datasets, UFC50 [39] and HMDB51 [40]; their visual samples are shown in Figure 3. The proposed method is implemented using Python (3.6 version) in Spyder integrated development environment.

A famous deep learning framework TensorFlow (2.5.0 version) with Keras backend and additional libraries including OpenCV, Scikit-image, and NumPy are used during experimentation. Furthermore, in the system configuration, Ubuntu operating system with GeForce RTX 2080-TI graphics card is used to accelerate the training process. The standard evaluation metrics such as Precision, Recall, F1-score, and Accuracy are used to evaluate the performance of the proposed method, as given in Table 4.

Table 5 shows the comparative results when compared to state-of-the-art models on UCF50 and HMDB51 datasets. In the subsequent sections, we define the datasets used in our experimentation along with discussion about our model's performance on the mentioned datasets.

Table 2: Different variants of ViT model used for image classification.

Model	Layers	Hidden size	MLP size	Heads	Params (M)
ViT-Base	12	768	3072	12	86
ViT-Large	24	1024	4096	16	307
ViT-Huge	32	1280	5120	16	632

The proposed method for features extraction is represented in bold text.

Table 3: The proposed LSTM network to capture long-range temporal information from video sequences.

Layer (type)	Output shape	No. of parameters
Input data	(None, 30, 1000)	0
LSTM	(None, 30, 128)	578048
LSTM	(None, 64)	49408
Dropout	(None, 64)	0
Batch normalization	(None, 64)	256
Activation	(None, 64)	0
Dense	(None, 64)	4160
Dense	(None, 51)	3315
Activation	(None, 51)	0

3.1. UCF50. UCF50 is a very popular HAR dataset consisting a total of 50 classes; all the video clips are collected from YouTube in ".avi" format. Each class in the video is divided into different groups that share common features; for example, in one group, a piano is played by a person four times but with a different viewpoint. Furthermore, it consists of a diverse collection of human activities due to high diversity in the camera motion, poses, and object appearances, viewpoints, clutter background, and different illumination in the surroundings. The performance of the proposed method is compared with different state-of-the art methods, for example, Handcrafted, LSTM, and Non-LSTM-based methods, as given in Table 5, where the last row shows the accuracy of the proposed method in percentage. Confusion matrix is given in Figure 4(a), where we have achieved 96.144% accuracy. Class-wise accuracy is shown in Figure 5, where the accuracies of the majority of classes are higher than 90%. For the comparative analysis, we have evaluated the performance of the proposed method with improved dense trajectories (IDT) hybrid approach [42], achieving 92.3% accuracy, while the LSTM-based method, the temporal optical flow with multi-layer LSTM [47], and lightweight CNN with DS-GRU [21] have achieved the second highest accuracies of 94.9% and 95.2%, respectively.

3.2. HMDB51. The HMDB51 dataset contains different varieties of videos related to human body movements such as facial interaction, object interaction with body, and human interaction for body movements. There are 6766 action video clips collected from different unique sources, all the video clips belong to 51 classes. Most of the activities are less than five seconds duration with each video frame resized into 224×224 dimensions for the training purpose. This dataset is very challenging because all the video clips are collected in challenging environments such as different illumination; four to six video clips of the same class and subjects are recorded in different pose and viewing orientations. Table 5 shows the comparative analysis of the

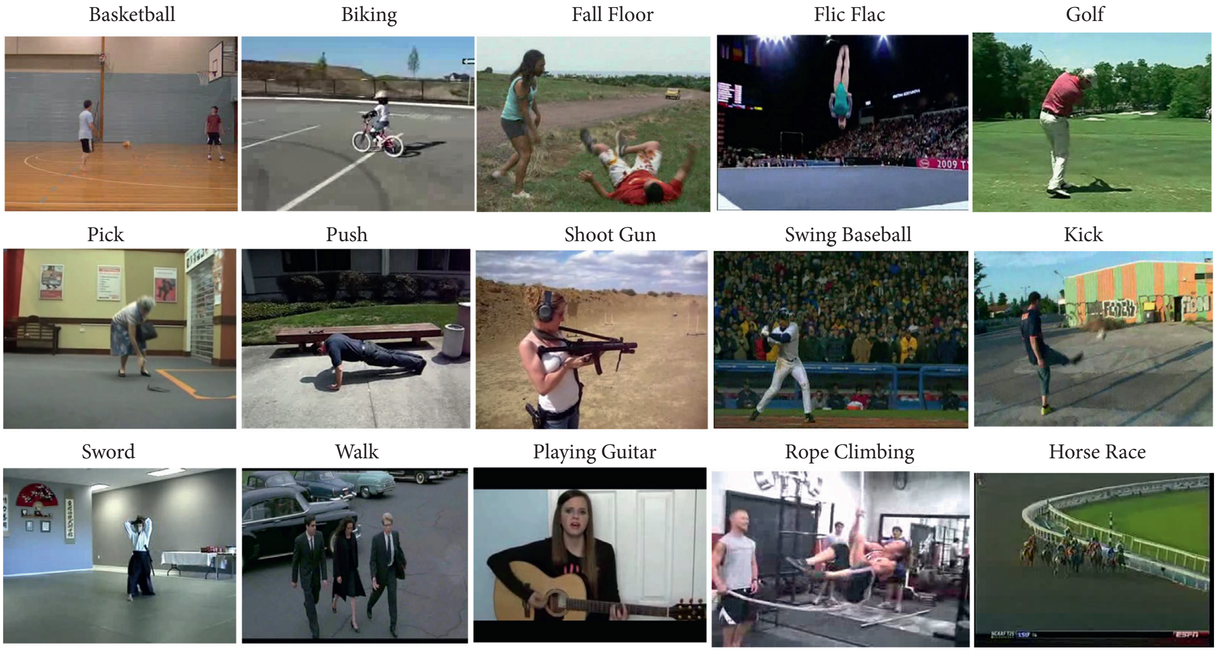


Figure 3: Sample action categories of UCF50 and HMDB51 datasets.

Table 4: Performance evaluation of the proposed method using Precision, Recall, and F1-score.

Dataset	Precision (%)	Recall (%)	F1-score (%)
UCF50	96.18655	96.14458	96.08283
HMDB51	76.49243	73.71429	73.51059

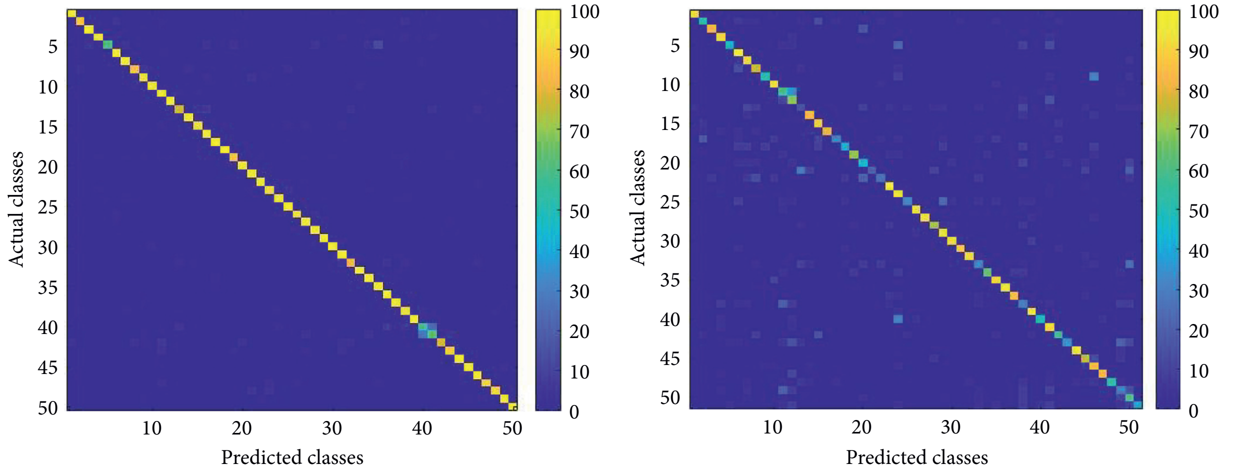
Table 5: Comparative analysis of the proposed method with traditional, deep LSTM, and non-LSTM-based techniques using HMDB51 and UCF50 datasets.

Domain	Technique	Accuracy (%)	
		HMDB51	UCF50
Handcrafted methods	Gradient boundary histogram + motion boundary descriptor [41]	62.2	--
	Improved dense trajectories (IDT) hybrid approach [42]	61.1	92.3
	Multiview super vector [43]	55.9	--
LSTM-based methods	Adoptive recurrent convolutional hybrid (ARCH) network [44]	58.2	--
	Lattice-LSTM [45]	66.2	--
	Relational LSTM [35]	71.4	--
	TS-LSTM and temporal inception [46]	69.0	--
	Temporal optical flow with multilayer LSTM [47]	72.2	94.9
	3D-CNNs and bidirectional hierarchical LSTM [48]	71.9	--
	CNN and DS-GRU [21]	72.3	95.2
Non-LSTM-based methods	Improved trajectory [49]	57.2	91.2
	Hierarchical clustering multitask learning [50]	51.4	93.2
The proposed method	ViT and multilayer LSTM	**73.714**	**96.144**

The methods represented by bold text show the highest performance in their respected categories.

proposed method with existing state-of-the-art methods such Handcrafted, LSTM, and Non-LSTM methods. The confusion matrix is shown in Figure 4(b), where the highest true positive value of each category is represented along in the diagonal; our proposed method achieved 73.714% accuracy. The class-wise accuracy is shown in Figure 6, where the horizontal axis represents numbers of classes, and the vertical axis shows the percentage accuracy of the corresponding category.

To summarise the existing literature, the highest accuracies of 62.2%, 72.3%, and 57.2% are achieved by Handcrafted, CNN and LSTM, and Non-LSTM-based methods, respectively. Meanwhile our proposed method improves the performance up to 1.414% on HMDB51 dataset when compared to these models, as given in Table 5. Similarly, the class-wise accuracy is shown in Figure 6, where our method achieved the best accuracy against rivals.



(a)

(b)

Figure 4: Confusion matrix of the proposed model. (a) UCF50 and (b) HMDB51 dataset.

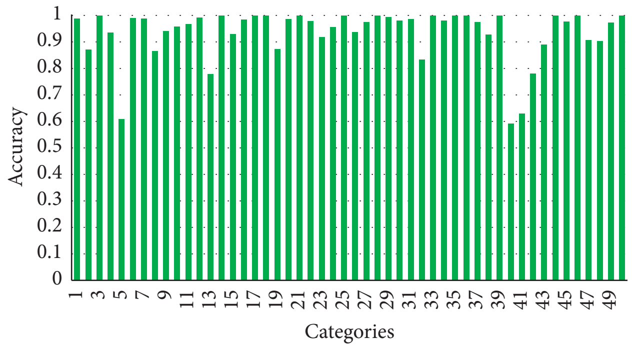


Figure 5: Class-wise accuracy of UCF50 dataset on the proposed ViT and multilayer LSTM model.

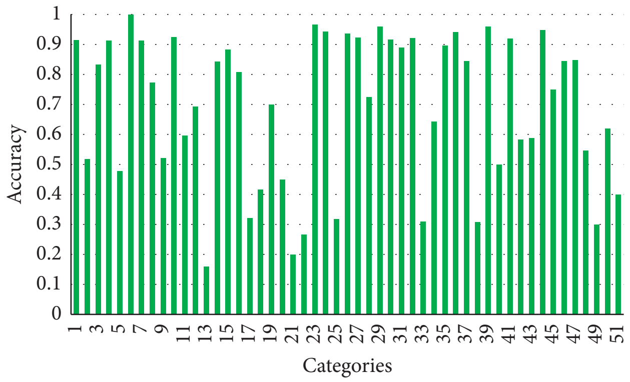


Figure 6: Class-wise accuracy of HMDB51 dataset on the proposed ViT and multilayer LSTM model.

4. Conclusions

An action is a sequence of multiple successive frames; thereby, both spatial and temporal features play an important role in accurate HAR. For this purpose, we have used pretrained ViT-Base-16 to extract the spatial features at predefined time stamps. These spatial features are fed to multilayered LSTM network to learn the long-range temporal dependencies. We have performed extensive experiments on two standard HAR datasets, UCF50 and IMDB51, and achieved recognition accuracies of 73.714% and 96.144%, respectively. However, in the proposed framework, there are few limitations that we will cover in the future research. For example, for HAR, we have used single-view camera that cannot provide full 360° coverage; in the future, we aim to use multiview data for efficient HAR. Moreover, the proposed framework is aimed to be transformed to an embedded platform to perform edge activity recognition. For efficient learning, two-stream networks with incremental learning strategy will be used to make it more intelligent to recognize complex actions in resource-constrained environments. Furthermore, in the future, we will combine different variants of ViT and different vision-based transformers models such as SWIN Transformer and ViViT for HAR. Our current system is flexible and can be helpful for adaptation in other domains such as emotion recognition, video summarization, and big data analytics.

Data Availability

The codes and related materials can be downloaded from https://github.com/Altaf-hucn/ViTLSTM-HAR.

Conflicts of Interest

The authors declare that there are no conflicts of interest.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (2019R1A2B5B01070067).

References

[1] A. Ullah, K. Muhammad, K. Haydarov, I. U. Haq, M. Lee, and S. W. Baik, "One-shot learning for surveillance anomaly recognition using siamese 3d cnn," in *Proceedings of the 2020 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, IEEE, Glasgow, UK, July 2020.

[2] M. Emad, M. Ishack, M. Ahmed, M. Osama, M. Salah, and G. Khoriba, "Early-anomaly prediction in surveillance cameras for security applications," in *Proceedings of the 2021 International Mobile, Intelligent, and Ubiquitous Computing Conference (MIUCC)*, pp. 124–128, IEEE, Cairo, Egypt, May 2021.

[3] W. Ullah, A. Ullah, T. Hussain et al., "Artificial Intelligence of Things-assisted two-stream neural network for anomaly detection in surveillance Big Video Data," *Future Generation Computer Systems*, vol. 129, pp. 286–297, 2021.

[4] A. Ullah, K. Muhammad, T. Hussain, S. W. Baik, and V. H. C. De Albuquerque, "Event-oriented 3D convolutional features selection and hash codes generation using PCA for video retrieval," *IEEE Access*, vol. 8, pp. 196529–196540, 2020.

[5] R. Jain, P. Jain, T. Kumar, and G. Dhiman, "Real time video summarizing using image semantic segmentation for CBVR," *Journal of Real-Time Image Processing*, vol. 18, no. 5, pp. 1827–1836, 2021.

[6] P. Kumar and S. Chauhan, "Retracted article: human activity recognition with deep learning: overview, challenges and possibilities," *CCF Transactions on Pervasive Computing and Interaction*, vol. 3, no. 3, p. 339, 2021.

[7] S. Ramasamy Ramamurthy and N. Roy, "Recent trends in machine learning for human activity recognition—a survey," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 8, no. 4, p. e1254, 2018.

[8] L. Onofri, P. Soda, M. Pechenizkiy, and G. Iannello, "A survey on using domain and contextual knowledge for human activity recognition in video streams," *Expert Systems with Applications*, vol. 63, pp. 97–111, 2016.

[9] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 1725–1732, IEEE, Columbus, OH, USA, June 2014.

[10] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *Computer Vision and Pattern Recognition*, vol. 6, p. 1556, 2014.

[11] S. U. Khan, I. U. Haq, S. Rho, S. W. Baik, and M. Y. Lee, "Cover the violence: a novel Deep-Learning-Based approach towards violence-detection in movies," *Applied Sciences*, vol. 9, no. 22, p. 4963, 2019.

[12] I. Serrano, O. Deniz, J. L. Espinosa-Aranda, and G. Bueno, "Fight recognition in video using hough forests and 2D convolutional neural network," *IEEE Transactions on Image Processing*, vol. 27, no. 10, pp. 4787–4797, 2018.

[13] R. Maqsood, U. I. Bajwa, G. Saleem, R. H. Raza, and M. W. Anwar, "Anomaly recognition from surveillance videos using 3D convolution neural network," *Multimedia Tools and Applications*, vol. 80, no. 12, pp. 18693–18716, 2021.

[14] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, pp. 4489–4497, IEEE, Santiago, Chile, Dec. 2015.

[15] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6299–6308, IEEE, Honolulu, HI, USA, July 2017.

[16] A. Hussain, K. Muhammad, H. Ullah et al., "Anomaly based camera prioritization in large scale surveillance networks," *Computers, Materials & Continua*, vol. 70, no. 2, pp. 2171–2190, 2022.

[17] X. Wang, L. Gao, P. Wang, X. Sun, and X. Liu, "Two-stream 3-d convnet fusion for action recognition in videos with arbitrary size and length," *IEEE Transactions on Multimedia*, vol. 20, no. 3, pp. 634–644, 2017.

[18] Z. Qiu, T. Yao, and T. Mei, "Learning spatio-temporal representation with pseudo-3d residual networks," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5533–5541, IEEE, Venice, Italy, Oct. 2017.

[19] Y. Zhou, X. Sun, Z.-J. Zha, and W. Zeng, "Mict: mixed 3d/2d convolutional tube for human action recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 449–458, IEEE, Salt Lake City, UT, USA, June 2018.

[20] M. Majd and R. Safabakhsh, "Correlational convolutional LSTM for human action recognition," *Neurocomputing*, vol. 396, pp. 224–229, 2020.

[21] A. Ullah, K. Muhammad, W. Ding, V. Palade, I. U. Haq, and S. W. Baik, "Efficient activity recognition using lightweight CNN and DS-GRU network for surveillance applications," *Applied Soft Computing*, vol. 103, p. 107102, 2021.

[22] K. Muhammad, F. Mustaqeem, A. Ullah et al., "Human action recognition using attention based LSTM network with dilated CNN features," *Future Generation Computer Systems*, vol. 125, pp. 820–830, 2021.

[23] W. Li, W. Nie, and Y. Su, "Human action recognition based on selected spatio-temporal features via bidirectional LSTM," *IEEE Access*, vol. 6, pp. 44211–44220, 2018.

[24] S. Ma, L. Sigal, and S. Sclaroff, "Learning activity progression in lstms for activity detection and early detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1942–1950, IEEE, Las Vegas, NV, USA, June 2016.

[25] A. Ullah, J. Ahmad, K. Muhammad, M. Sajjad, and S. W. Baik, "Action recognition in video sequences using deep bi-directional LSTM with CNN features," *IEEE access*, vol. 6, pp. 1155–1166, 2017.

[26] W. Ullah, A. Ullah, T. Hussain, Z. A. Khan, and S. W. Baik, "An efficient anomaly recognition framework using an attention residual LSTM in surveillance videos," *Sensors*, vol. 21, no. 8, p. 2811, 2021.

[27] T. Hussain, K. Muhammad, A. Ullah et al., "Multiview summarization and activity recognition meet edge computing in IoT environments," *IEEE Internet of Things Journal*, vol. 8, no. 12, pp. 9634–9644, 2020.

[28] M. Ott, S. Edunov, D. Grangier, and M. Auli, "Scaling neural machine translation," in *Proceedings of the Third Conference on Machine Translation: Research Papers*, pp. 1–19, ACL, Brussels, Belgium, Oct. 2018.

[29] A. Fan, S. Bhosale, H. Schwenk et al., "Beyond English-centric multilingual machine translation," *Journal of Machine Learning Research*, vol. 22, no. 107, pp. 1–48, 2021.

[30] T. B. Brown, B. Mann, N. Ryder et al., "Language models are few-shot learners," in *Proceedings of the 1st Workshop on Multilingual Representation Learning*, Oct. 2020.

[31] Z. Dai, Z. Yang, Y. Yang, J. Carbonell, Q. V. Le, and R. Salakhutdinov, "Transformer-xl: attentive language models beyond a fixed-length context," in *Proceedings of the 57th*

Annual Meeting of the Association for Computational Linguistics, ACL, Brussels, Belgium, Oct. 2019.

[32] A. Vaswani et al., "Attention is all you need," in *Proceedings of the Advances in neural information processing systems*, pp. 5998–6008, ACL, Brussels, Belgium, Oct. 2017.

[33] A. Dosovitskiy et al., "An image is worth 16x16 words: transformers for image recognition at scale," *Computer Vision and Pattern Recognition*, vol. 2, 2020.

[34] S. Stylianou-Nikolaidou, I. Vernikos, E. Mathe, E. Spyrou, and P. Mylonas, "A novel CNN-LSTM hybrid architecture for the recognition of human activities," *Proceedings of the International Neural Networks Society*, Springer, pp. 121–132, July 2021.

[35] Z. Chen, B. Ramachandra, T. Wu, and R. R. Vatsavai, "Relational long short-term memory for video action recognition," *Computer Vision and Pattern Recognition*, vol. 1, 2018.

[36] F. U. M. Ullah, M. S. Obaidat, K. Muhammad et al., "An intelligent system for complex violence pattern analysis and detection," *International Journal of Intelligent Systems*, 2021.

[37] J. G. Zilly, R. K. Srivastava, J. Koutnık, and J. Schmidhuber, "Recurrent highway networks," in *Proceedings of the International Conference on Machine Learning*, pp. 4189–4198, PMLR, New York City, NY, USA, June 2017.

[38] N. Liu, N. Zhang, K. Wan, L. Shao, and J. Han, "Visual saliency transformer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4722–4732, IEEE, Barcelona, Spain, OCT.2021.

[39] K. K. Reddy and M. Shah, "Recognizing 50 human action categories of web videos," *Machine Vision and Applications*, vol. 24, no. 5, pp. 971–981, 2013.

[40] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "HMDB: a large video database for human motion recognition," in *Proceedings of the 2011 International conference on computer vision*, pp. 2556–2563, IEEE, Barcelona, Spain, Nov. 2011.

[41] F. Shi, R. Laganiere, and E. Petriu, "Gradient boundary histograms for action recognition," in *Proceedings of the 2015 IEEE Winter Conference on Applications of Computer Vision*, pp. 1107–1114, IEEE, Waikoloa, HI, USA, Jan. 2015.

[42] X. Peng, L. Wang, X. Wang, and Y. Qiao, "Bag of visual words and fusion methods for action recognition: comprehensive study and good practice," *Computer Vision and Image Understanding*, vol. 150, pp. 109–125, 2016.

[43] Z. Cai, L. Wang, X. Peng, and Y. Qiao, "Multi-view super vector for action recognition," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 596–603, Columbus, OH, USA, 2014.

[44] M. Xin, H. Zhang, H. Wang, M. Sun, and D. Yuan, "Arch: adaptive recurrent-convolutional hybrid networks for long-term action recognition," *Neurocomputing*, vol. 178, pp. 87–102, 2016.

[45] L. Sun, K. Jia, K. Chen, D.-Y. Yeung, B. E. Shi, and S. Savarese, "Lattice long short-term memory for human action recognition," in *Proceedings of the IEEE international conference on computer vision*, pp. 2147–2156, IEEE, Venice, Italy, Oct. 2017.

[46] C.-Y. Ma, M.-H. Chen, Z. Kira, and G. AlRegib, "TS-LSTM and temporal-inception: exploiting spatiotemporal dynamics for activity recognition," *Signal Processing: Image Communication*, vol. 71, pp. 76–87, 2019.

[47] A. Ullah, K. Muhammad, J. Del Ser, S. W. Baik, and V. H. C. de Albuquerque, "Activity recognition using temporal optical flow convolutional features and multilayer LSTM," *IEEE Transactions on Industrial Electronics*, vol. 66, no. 12, pp. 9692–9702, 2018.

[48] H. Yang, J. Zhang, S. Li, and T. Luo, "Bi-direction hierarchical LSTM with spatial-temporal attention for action recognition," *Journal of Intelligent and Fuzzy Systems*, vol. 36, no. 1, pp. 775–786, 2019.

[49] H. Wang and C. Schmid, "Action recognition with improved trajectories," in *Proceedings of the IEEE international conference on computer vision*, pp. 3551–3558, IEEE, Sydney, NSW, Australia, Dec. 2013.

[50] A.-A. Liu, Y.-T. Su, W.-Z. Nie, and M. Kankanhalli, "Hierarchical clustering multi-task learning for joint human action grouping and recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 1, pp. 102–114, 2016.