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

Copyright © 2022 Altaf Hussain et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

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 3D 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 of 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 \( x_p \in \mathbb{R}^{N \times (P^2 \times C)} \). 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 \(16 \times 16\) or \(32 \times 32\), 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 reorder the spatial information as the original image. The result of embedded patches with token \(Z_0\) is mathematically represented in (1).

\[
Z_0 = [v_{class}; x_1 E; x_2 E; \ldots x_N E] + E_{pos}, E \in \mathbb{R}^{(P^2 \times C \times d)}, E \in \mathbb{R}^{(N+1) \times d}.
\]

(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 multhead 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.

\[
z_l = MSA(LN(z_{l-1})) + z_{l-1}, l = 1 \ldots L,
\]

(2)

\[
z_l = MLP(LN(z_0^l)) + z_p^l, l = 1 \ldots L.
\]

(3)

In the last layer of the encoder, the first element \(z_0^L\) in the sequence is passed to external head classifier for predicting the class label.

\[
y = LN(z_0^L).
\]

(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 $U^Q_{K,V}$; a single SA is graphically represented in Figure 2(d), while the mathematical formulation is given in equation (5):

$$[Q, K, V] = zU^Q_{K,V}, \quad U^Q_{K,V} \in \mathbb{R}^{d_x \times d_k}.$$ (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):

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

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):

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

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

<table>
<thead>
<tr>
<th>Variables/symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\Delta t$</td>
<td>Input over time (t)</td>
</tr>
<tr>
<td>$\partial$</td>
<td>Sigmoid activation function</td>
</tr>
<tr>
<td>$w$</td>
<td>Weights</td>
</tr>
<tr>
<td>$b$</td>
<td>Bias terms</td>
</tr>
<tr>
<td>$i_t$</td>
<td>Input gate</td>
</tr>
<tr>
<td>$f_t$</td>
<td>Forget gate</td>
</tr>
<tr>
<td>$o_t$</td>
<td>Output gate</td>
</tr>
<tr>
<td>$tanh$</td>
<td>Tan (h) activation function</td>
</tr>
<tr>
<td>SoftMax</td>
<td>Activation for the final classification</td>
</tr>
<tr>
<td>$N$</td>
<td>Numbers of classes</td>
</tr>
</tbody>
</table>

#### 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 $\Delta t$ 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.

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

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

$$o_t = \partial (w_o[\Delta t + S_{t-1}] + b_o), \quad (10)$$

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

$$C_t = C_{t-1} \cdot f_t + R \cdot i_t, \quad (12)$$

$$S_t = \tanh (C_t) \cdot O_t, \quad (13)$$

$$\text{Prediction}_{\text{state}} = \text{soft max}(s_tN). \quad (14)$$

Herein, the term $\Delta t$ represents the input over time and sigmoid activation function is represented by $\partial$. 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 UCF500 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
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. For experimentation along with discussion about our model’s performance on the mentioned datasets.

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

---

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

<table>
<thead>
<tr>
<th>Model</th>
<th>Layers</th>
<th>Hidden size</th>
<th>MLP size</th>
<th>Heads</th>
<th>Params (M)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ViT-Base</td>
<td>12</td>
<td>768</td>
<td>3072</td>
<td>12</td>
<td>86</td>
</tr>
<tr>
<td>ViT-Large</td>
<td>24</td>
<td>1024</td>
<td>4096</td>
<td>16</td>
<td>307</td>
</tr>
<tr>
<td>ViT-Huge</td>
<td>32</td>
<td>1280</td>
<td>5120</td>
<td>16</td>
<td>632</td>
</tr>
</tbody>
</table>

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.**

<table>
<thead>
<tr>
<th>Layer (type)</th>
<th>Output shape</th>
<th>No. of parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input data</td>
<td>(None, 30, 1000)</td>
<td>0</td>
</tr>
<tr>
<td>LSTM</td>
<td>(None, 30, 128)</td>
<td>578048</td>
</tr>
<tr>
<td>LSTM</td>
<td>(None, 64)</td>
<td>49408</td>
</tr>
<tr>
<td>Dropout</td>
<td>(None, 64)</td>
<td>0</td>
</tr>
<tr>
<td>Batch normalization</td>
<td>(None, 64)</td>
<td>256</td>
</tr>
<tr>
<td>Activation</td>
<td>(None, 64)</td>
<td>0</td>
</tr>
<tr>
<td>Dense</td>
<td>(None, 64)</td>
<td>4160</td>
</tr>
<tr>
<td>Dense</td>
<td>(None, 51)</td>
<td>3315</td>
</tr>
<tr>
<td>Activation</td>
<td>(None, 51)</td>
<td>0</td>
</tr>
</tbody>
</table>

---
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.
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


