Abnormal behavior recognition using SRU with attention mechanism

Nian Chi Tay [a,1], Tee Connie [b,2,*], Thian Song Ong [b,3], Andrew Beng Jin Teoh [b,4], Pin Shen Teh [c,5]

[a] Faculty of Information Science and Technology, Multimedia University, Jalan Ayer Keroh Lama, 75450 Bukit Beruang, Malacca, Malaysia
[b] School of Electrical and Electronic Engineering, College of Engineering, Yonsei University, 50 Yonsei-ro, Sinchon-dong, Seoul, South Korea
[c] Department of Operations, Technology, Events and Hospitality Management, Manchester Metropolitan University, Oxford Road, Manchester, M15 6BH, United Kingdom
[1] nianchi.tay95@gmail.com; [2] tee.connie@mmu.edu.my; [3] tsong@mmu.edu.my; [4] bjteoh@yonsei.ac.kr; [5] p.teh@mmu.ac.uk
* corresponding author

ARTICLE INFO

Article history
Received October 20, 2023
Revised February 6, 2024
Accepted February 23, 2024
Available online May 31, 2024

Keywords
Abnormal behavior recognition
Simple recurrent unit
Attention mechanism
Long short-term memory

ABSTRACT

In response to the critical need for enhanced public safety measures, this study introduces an advanced intelligent surveillance system designed to autonomously detect abnormal behaviors within public spaces. Leveraging the computational efficiency and accuracy of a Simple Recurrent Unit (SRU) integrated with an attention mechanism, this research delivers a novel approach towards understanding and interpreting human interactions in real-time video footage. Distinctively, the model specializes in identifying two primary categories of abnormal behavior: aggressive two-person interactions such as physical confrontations and collective crowd dynamics, characterized by sudden dispersal patterns indicative of distress or danger. The incorporation of Attention mechanism precisely targets critical elements of behavior, thereby enhancing the model's focus and interpretative clarity. Empirical validation across five benchmark datasets reveals that our model not only outperforms traditional Long Short-Term Memory (LSTM) frameworks in terms of speed by a factor of 1.5 but also demonstrates superior accuracy in abnormal behavior recognition. These findings not only underscore the model's potential in preempting potential safety threats but also mark a significant advancement in the application of deep learning technologies for public security infrastructures. This research contributes to the broader discourse on public safety, offering actionable insights and robust technological solutions to enhance surveillance efficacy and response mechanisms in critical public domains.

1. Introduction

The rising trend of security technology has been a hot topic for society nowadays [1]–[5]. Every now and then, there are news involving crimes like robbery, terrorism, or fighting cases everywhere [6]–[9]. Hence, many surveillance devices have been installed to ensure public safety in places such as banks, schools, shops, and subway stations [10], [11]. However, conventional surveillance cameras may not be effective enough as laborious efforts are required [12] to keep on monitoring the cameras for any abnormal behaviors. To improve efficiency, a solution that supports autonomous recognition is needed to effectively recognize abnormal behavior [13]–[21].

Abnormal behavior, by definition, encompasses actions that deviate from the norm within a given context [22], such as aggressive confrontations between individuals or panic-driven crowd movements. These behaviors, particularly in public settings like malls, schools, or transportation hubs, often presage

incidents of violence or emergency situations. Thus, the automated detection and timely recognition of such behaviors can significantly enhance response mechanisms and potentially mitigate adverse outcomes.

Building on the foundation laid by [23], who introduced the concept of neural machine attention in their seminal work, this paper proposes an innovative abnormal behavior recognition method that incorporates a Simple Recurrent Unit (SRU) with an attention mechanism. The attention mechanism's introduction to the field of neural networks marked a pivotal shift towards models capable of selectively focusing on parts of the input data, enhancing the model's ability to interpret complex data patterns. Our approach extends this paradigm by implementing feature extraction through Convolutional Neural Networks (CNN) followed by the strategic application of attention mechanisms to distill critical features indicative of abnormal behavior.

The urgency to address the limitations of existing models is underscored by the evolving complexity of surveillance environments and the dynamic nature of abnormal behaviors. Previous efforts have predominantly focused on either two-person interactions or crowd dynamics in isolation, with limited success in real-time application scenarios due to high false-positive rates and computational inefficiency. This research seeks to bridge these gaps by presenting a model that not only outperforms traditional LSTM-based approaches in speed but also introduces a nuanced understanding of abnormal behavior recognition through the integration of SRU and attention mechanisms. Three contributions from this study are summarized as follows:

- Visualization of how the model "sees" and interprets the results based on where the attention mechanism focuses on.
- To the best of our knowledge, SRU has not yet been used in abnormal behavior recognition. The findings in our study show that SRU is able to obtain higher recognition accuracy and faster recognition speed than conventional LSTM.
- The performance of the model was validated on five benchmark datasets namely, (a) two-person interactions: Hockey Fight Dataset (HCF) [24], Peliculas Dataset (PEL) [25], UTI Dataset (UTI) [26], and (b) crowd-based interactions: UMN Dataset (UMN) [27] and Web Dataset (WEB) [28].

2. Related Works

Many studies have been conducted to design a robust system in abnormal behavior recognition and the state-of-the-art deep learning method is the most popular among these studies. Our previous work [29] focused on the study of CNN and LSTM in recognizing abnormal behavior, however, the time consumption during the training phase was one of the limitations that we found. This paper is inspired by the work of Lei et al. [30] which used SRU in multiple Natural Language Processing (NLP) tasks. The significance of SRU introduced by the authors is that it offers high parallelization and achieves accurate results for sequence modeling problems. The authors proved that SRU's performance can be competed with a feed-forward network in terms of computational speed which is 5-9 times faster compared to cuDNN-optimised LSTM. This can be realized by parallelizing all the computations across the hidden states and timesteps. Different from the normal recurrent network, SRU has fully utilized the advantage of GPUs. Another work by Ko and Sim [31] demonstrated good performance using deep convolutional networks. Transfer learning was carried out on LSTM and Kalman filter was adopted for the detection of close interactions among human subjects. Arifoglu and Bouchachia [32] looked into the abnormal behavior of dementia people by employing Recurrent Neural Network (RNN) techniques such as Gated Recurrent Unit RNNs (GRU), Vanilla RNNs (VRNN) and LSTM. Features like change-point and last-fired representations were extracted and the results showed that RNN is suitable in the field of activity recognition.

Sultani et al. [12] introduced anomaly detection in real-world surveillance videos using multiple instance learning (MIL) and ranking methods. Normal and abnormal videos were represented as negative and positive bags respectively. The videos were then divided into temporal segments and stored as

instances in the bags. The instances were passed to the pre-trained 3D ConvNet to extract the features as they were proved to be computational efficient [12]. A fully connected neural network trained the features by implementing ranking loss function to compute a ranking loss between the highest score instances in both positive and negative bags. The authors also proposed sparsity and temporal constraints to further enhance the ranking loss function in localizing anomaly. Fan et al. [33] introduced an early event detection system using surveillance videos and dynamic images as the input to the Deep ConvNet. The authors outlined the experiments based on two categories of incidents: human falling and fighting scenes. Dynamics images were converted from both normal and abnormal events video clips. Transfer learning was carried out using ImageNet dataset and followed by a fine-tuned VGG16 model. The dynamic images with static background and "shadow-black" regions formed around the human showed that falling or fighting action was performed.

In various studies, the attention mechanism has been employed to concentrate on specific segments of the input within neural networks. Sharma et al. [34] utilized visual attention to selectively focus on parts of video frames, aggregating predictions from each frame to determine the final label, thus achieving superior accuracy over traditional max or average pooling methods by dynamically pooling convolutional features. They utilized the feature slice from each timestep as input for a three-layer LSTM network. Chen et al. [35] introduced the Attentional ConvLSTM (AC-LSTM), which applied the attention mechanism across both high-level and low-level features, with the ConvLSTM's hidden states conducting multi-box regression and classification to efficiently manage temporal memory by filtering out irrelevant information. Karpathy et al. [36] employed CNNs for extensive video classification on the UCF-101 dataset, designing attention within the fovea stream to focus on the frame's central region.

Differently, Xu et al. [37] integrated both soft deterministic and hard stochastic attention mechanisms within LSTM to frame-wise describe videos, yielding results that notably surpassed human evaluations. Jaderberg et al. [38] introduced a soft attention variant, the Spatial Transformer, which, when inserted between CNN layers, set a new benchmark in performance, particularly demonstrated by its efficacy in recognizing street view house numbers. These pivotal studies in video recognition are concisely summarized in Table 1, showcasing the diverse applications and outcomes of integrating attention mechanisms into video analysis.

Table 1. Summary of Existing Work

Authors	Methodology	Dataset(s)	Dataset(s) Descriptions	Highest Accuracy (%)
Ko and Sim [31]	YOLO v2 + VGG16 + LSTM with Kalman filter	UT-Interaction	Two-person interactions	95
Arifoglu and Bouchachia [32]	VRNN, LSTM and GRU	Van Kasteren	Dementia people	96.7
Chen et al. [35]	AC-LSTM	ImageNet VID, 2DMOTI5	Real-world object detection	65.43
Karpathy et al. [36]	CNN with fovea stream	UCF-101	Human activity recognition	65.4
Sharma et al. [34]	Visual attention with LSTM	UCF-11, HMDB-51, Hollywood2	Human activity recognition	43.9

While several studies have explored the use of deep learning models for abnormal behavior recognition, our approach distinguishes itself through the novel application of SRU and attention mechanisms. Unlike previous models that have struggled with the dual challenges of computational efficiency and accuracy, our model demonstrates a balanced improvement in both areas. This comparison underscores the potential of our methodology to set a new benchmark in the field, offering a viable solution to the longstanding challenges of abnormal behavior detection in surveillance footage.

3. Method

Our proposed approach consists of two different methods, (i) attention mechanism and (ii) SRU. The following section describes each of this method in further detail and the last part entails the combination of these two methods to form the abnormal behavior recognition model.

3.1. Attention Mechanism

The attention mechanism, traditionally utilized in encoder-decoder models for tasks like video description and machine translation, is also beneficial for enhancing abnormal behavior recognition systems. By integrating it within the feature extraction layer, it's possible to meticulously analyze video sequence images' salient regions, significantly boosting system performance. In our study, we have incorporated the attention mechanism into video recognition, implementing it directly within Convolutional Neural Networks (CNNs) for model training. Among the attention mechanism variants, soft and hard attention [39], we opted for the soft attention algorithm as described by [37]. This choice was driven by its deterministic approach, ensuring compatibility with standard backpropagation algorithms for efficient model learning:

$$a = f\emptyset(x) \tag{1}$$

where a and x represent an attention and input vector respectively, while $f\emptyset(x)$ represents the attention network with parameter $\emptyset$. The attention glimpse, g, is represented in equation below:

$$g = a \odot z \tag{2}$$

where z represents the output of subsequent neural network $f\emptyset(x)$.

A mask with range zero to one is multiplied with the features in soft attention. The importance of attention mechanism is its simplicity and compacity. The difference between attention mechanism and conventional neural network is that the latter makes use of series of matrix multiplications and element-wise non-linearities. This leads to the fact that the interaction among features vectors are performed using repeated addition only. However in attention mechanism, the features are multiplied among each other using the soft mask computed. With the universal function approximators, the neural networks are limited to certain number of hidden units due to its vast amount of computation during features interactions. The attention mechanism facilitates the computation by introducing multiplicative interactions instead of repeated addition interactions. As such, the capabilities of neural networks are expanded as more complicated functions are able to be approximated to enable the ability of focusing on salient parts of the input.

3.2. Simple Recurrent Unit (SRU)

The primary objective behind adopting the SRU is to decrease the network training time while preserving its accuracy. The operational procedure of SRU aligns closely with that of LSTM and other gated recurrent networks, with the distinct difference lying in the calculation method utilized within the sigmoid gate. Fig. 1 illustrates the architectural differences between a conventional RNN and an SRU.

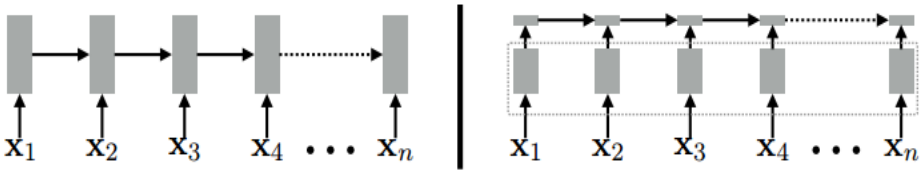

Fig. 1. Comparison of Architecture between RNN (left) and SRU (right) [40].

Fig. 1 demonstrates that the outputs at each timestep within the SRU are processed in parallel. The distinct features of the SRU can be categorized into two primary components: (i) light recurrence and (ii) the highway network. The concept of **Light Recurrence** is detailed through the following equations:

$$f_t = \sigma\left(W_{fX_t} + V_f \odot C_{t-1} + b_f\right) \quad (3)$$

$$c_t = f_t \odot C_{t-1} + (1 - - f_t) \odot (W_{X_t}) \quad (4)$$

where W and W_f are the weights and v_f and b_f represent the parameter vectors. The model first reads the input x_r and decides which information to discard in the forget gate. Light recurrence also performs computation for the state C_t Similar to LSTM and other gated recurrent networks, current state C_t still depends on the previous state, C_{t-1} and based on W_{Xt} according to the forget gate, f_t. The main difference between SRU and LSTM is how C_{t-1}being multiplied during the computation. The typical LSTM uses normal multiplication between parameter vector value of forget gate, v_f and C_{t-1} as stated in the equation below:

$$f_t = \sigma(W_{fX_t} + V_f C_{t-1} + b_f \quad (5)$$

However, the problem occurred with this kind of multiplication is that parallelization is hard to achieve as each state vector needs to wait until previous state, C_{t-1} has done its computation. Hence, Lei et al. [30] proposed point-wise multiplication to facilitate the computation so that each vector state becomes independent and can be parallelized.

Highway Network is the second component in SRU. While the first component is to enhance the parallelization of the network, highway network [41] is adopted to facilitate gradient propagation. This can be denoted by the equations below:

$$r_t = \sigma(W_{rX_t} + V_r \odot C_{t-1} + b_r \quad (6)$$

$$h_t = r_t \odot C_{t-1} + (1 - - r_t) \odot (X_t) \quad (7)$$

where r_t is the reset gate and W is the weight parameter. The reset gate is combined with the state c_t produced by the light recurrence. In order to propagate the gradient directly to the previous layer, a skip connection is proposed where $(1 - - r_t) \odot X_t$ to improve the speed of computation in the network training.

3.3. SRU with Attention Mechanism

Our proposed model combines Simple Recurrent Unit (SRU) with attention mechanism to enhance abnormal behavior recognition in surveillance videos. This section delves into the intricate details of our architecture, specifically highlighting the role of the attention mechanism.

The architecture is based on an Encoder-Decoder framework, where the encoder processes the input sequence to create a context-rich representation. The decoder then generates the output sequence, relying on the context provided by the encoder. Our model's encoder consists of convolutional layers that extract salient features from input video frames. These features serve as the basis for recognizing abnormal behaviors, such as unexpected gatherings or violent actions. The attention mechanism is integrated between the encoder and decoder, focusing on the most relevant features extracted by the encoder. Unlike traditional models where the entire context is passed uniformly to the decoder, our attention mechanism dynamically weighs the importance of different parts of the input sequence. This process ensures that the model pays more attention to the segments of the video where abnormal behavior is more likely to be present. We employed a soft attention algorithm that allows the model to focus on specific parts of the input sequence without ignoring the rest. This algorithm computes attention weights, which are then used to create a weighted sum of the encoder's outputs, forming a context vector. The context vector is then fed into the decoder, guiding the generation of the output sequence. In the attention layer, an element-wise multiplication is performed between the encoder's output and the computed attention weights. This step highlights the features of utmost importance, ensuring that the decoder's focus is directed towards the most relevant information for abnormal behavior detection.

The context vector, enriched with focused attention, significantly influences the decoder's processing. By providing a dynamically weighted context, the decoder can make more informed predictions, enhancing the accuracy of abnormal behavior recognition. To optimize computational efficiency and reduce costs, we processed the input images by converting them to grayscale and resizing them to a more manageable resolution of 128x128 pixels. During the preprocessing phase, we employed adaptive histogram equalization to improve the local contrast within the images, thereby enhancing the edge definition across various image regions. The network architecture depicted in Fig. 2 comprises three key components: an input layer for processing the images, convolutional layers enhanced with an attention mechanism to focus on pertinent features, and an SRU layer tasked with interpreting the features extracted by the CNN.

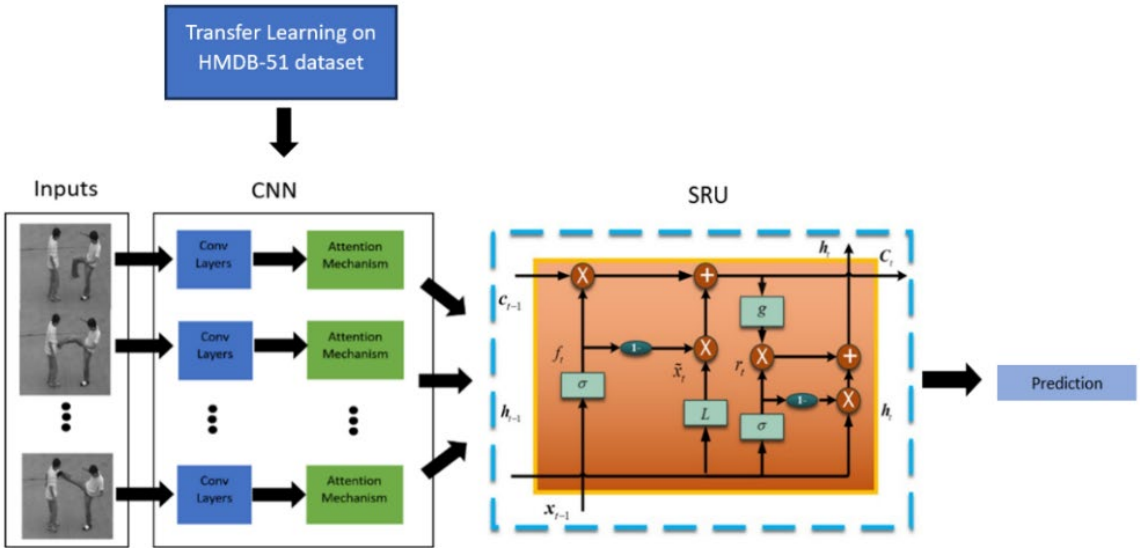

Fig. 2. Framework of SRU with attention mechanism

First, the input layer takes in input images and passes them to convolutional layers for feature extraction. In the feature extraction phase, low-level features such as human edges to high-level features like high dimensional body parts are extracted and activation map is formed. For the first convolutional layer, 64 filters of size 3x3 are used. We test the number of filters empirically by increasing its number as the network goes deeper in order to find the optimum number of filters. The model includes zero padding to make sure that both output and input lengths are the same for the purpose of maintaining temporal order of data. To make the training smoother, all negative activations are changed to zero by applying the activation function, Rectified Linear Units (ReLU) after the convolutional layer. A max pooling layer of strides 2 pixels is added to reduce the feature map's dimensionality while retaining most of the input's important information. As the network gets deeper, we increase the convolutional layer's filter number to 128. Another ReLU activation function and max pooling layer are added to the network. Finally, attention layer is applied to the network to carry out element-wise multiplication between the previous input and the output.

The model repeats the same process of CNN and attention mechanism for the next input image until the inputs have reached the stated timestep. SRU will then analyze the temporal information from the features extracted by the previous layer. In SRU layer, the number of hidden cells (256, 512, 1024) are tested. The optimization algorithm used is Adam optimizer which uses an adaptive learning rate method according to the weights of the neural network which is computational efficient during training. The network includes gradient clipping of 1 to prevent an exploding gradient effect. The model details and parameters used are summarized in Fig. 3.

Parameters	Conv1, Pool1, ReLU1	Conv2, Pool2, ReLU2
Conv. Filters	64	128
Conv. Filter Size	3x3	3x3
Max Pooling Stride	2	2

Parameters	Attention Mechanism Layer
No. of Filters	128
Layer	Dense
Activation Function	Relu

Parameters	SRU Layer(s)
Cell Units	256/512/1024

Parameters	Dense Layer
Output units	2
Activation Function	Softmax

Fig. 3. Summary of model layers

4. Results and Discussion

The problem incurred in this work is the limited amount of data. Hence, a data augmentation technique is adopted to ensure effective deep learning. The original data are augmented and reproduced using methods such as horizontal flip, height shift range of 0.1 to 0.2, and width shift range of -5 to 10. The datasets details are shown in Table 2. Transfer learning was carried out by pre-training the HMDB-51 Dataset [42] which consists of 51 classes of human activities like punching, kicking someone, and gun-shooting. The pre-trained network achieved 99% accuracy and were used by the five benchmark datasets mainly HCF, PEL, UTI, UMN and WEB as the base model.

Table 2. Datasets details after data augmentation

Dataset	Videos				Images				Total			
	Training		*Testing*		*Training*		*Testing*		*Videos*		*Images*	
	*Ab **	*Nor **	*Ab*	*Nor*	*Ab*	*Nor*	*Ab*	*Nor*	*Train*	*Test*	*Train*	*Test*
Hockey Fight Dataset (HCF)	700	700	300	300	21000	21000	9000	9000	1400	600	42000	18000
Peliculas Dataset (PEL)	700	707	300	303	21000	21210	9000	9090	1407	603	42210	18090
UT-Interactions Dataset (UTI)	756	731	324	313	22680	21930	9720	9390	1487	637	44610	19110
UMN Dataset (UMN)	718	4158	308	1782	21540	124740	9240	53460	4887	2079	146610	62370
Web Dataset (Web)	700	1050	300	450	21000	9000	31500	13500	1800	750	54000	22500

[a] Ab*: Abnormal; Nm*: Normal

The experiments are carried out in two parts: (i) Experiment 1: for two-person interactions [24]–[26] and (ii) Experiment 2: for crowd-based interactions [27], [28]. Both experiments were carried out using Python 3.5.2 Keras with the aid of GPU-Tensorflow 1.11.0 (CUDA version 9.0) on a workstation equipped with NVIDIA GeForce GTX 1080. Various libraries are adopted such as NumPy, scikit-learn, and Matplotlib. The parameters for both Experiments 1 and 2 are listed in Table 3.

Table 3. Summary of parameters for Experiments 1 and 2

Parameters	Values
Number of epochs	10, 100
Hidden units of SRU cells	256, 512, 1024
Learning rate	0.001
Adam optimizer	beta1=0.9, beta2=0.998

The hidden units of SRU cells were tested with 256, 512 and 1024 units [31] to determine which value holds better effect on the model performance. AMSGrad is applied to slow down the learning rate from decaying. The loss function employed is binary cross-entropy as the output category is only two, either normal behavior or abnormal behavior.

4.1. Experiment 1

This experiment involves two-person interactions such as kicking, punching, pushing, and fighting by using HCF, PEL and UTI datasets. The purpose of this experiment is to investigate the robustness of the model in identifying abnormal behaviors between two-person. The attention mechanism is applied at the feature extraction stage to determine where are the focus of the model when extracting the salient features from the images. Fig. 4 - Fig. 6 show the screenshots of attention in a given set of video frames from HCF, PEL, and UTI datasets respectively. The first row represents the feature map of SRU without attention mechanism and the second row shows the feature map of SRU with attention mechanism.

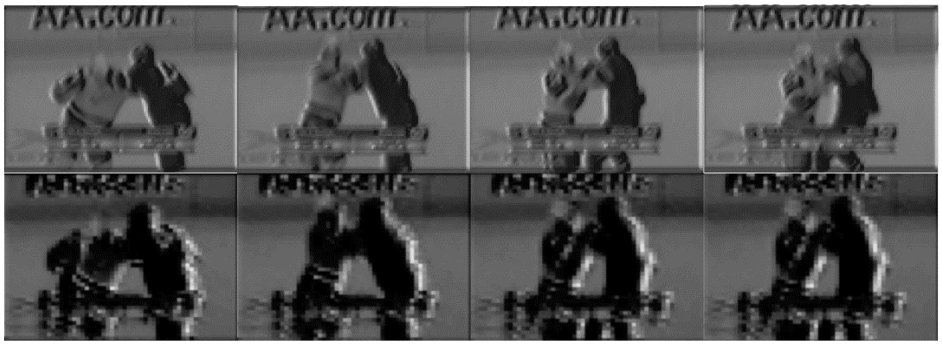

Fig. 4. Visualization of HCF dataset

Fig. 5. Visualization of PEL dataset

The attention mechanism highlights the hockey players' bodylines like the right player's back and the left player's hand. It can be seen that without an attention mechanism, the feature map is sparse and the focus cannot be visualized clearly. Fig. 5 mentions the observation of the model applied to the PEL.

As can be seen, the SRU model without an attention mechanism does not focus much on the human subjects instead, it highlights the background of the image. On the other hand, the second row represents SRU with attention mechanism and focuses on the top right corner of the image and also the legs of the two fighters.

This is again the case for UTI as shown in Fig. 6. The bodylines of two subjects are focused by the attention mechanism and the highlighted parts such as legs proves vital when recognizing a kicking action. Without attention mechanism, the focus of the model becomes ambiguous as shown in the first row, which only highlights the hand of the person on the left.

Fig. 6. Visualization of UTI dataset

From the visualization results shown in the three figures, it is clear that the attention mechanism focuses on human's bodylines when extracting the features from the images. The highlighted features from the attention mechanism are very important for abnormal behavior recognition as human actions involve body motions. Therefore, the body lines of human are essential when recognizing actions. For example, the hand stretching pattern indicates that the human is performing actions related to hands such as punching, pushing and so on.

The accuracy is computed based on the Keras accuracy calculation for binary classification by taking 0.5 as the threshold for identifying the category as of whether it is normal behavior or abnormal behavior based on the mean of the predictions on the true labels. The accuracies of the model on 100 number of epochs, cell units, and layers based on 0.001 learning rate are summarized in Table 4, Table 5, and Table 6 for SRU with and without attention mechanism, and LSTM with attention mechanism.

Table 4. Accuracy and computational time of HCF with 100 epochs

	Cell Units = 256			**Cell Units = 512**			**Cell Units = 1024**		
	SRU with AM	***SRU w/o AM***	***LSTM with AM***	***SRU with AM***	***SRU w/o AM***	***LSTM with AM***	***SRU with AM***	***SRU w/o AM***	***LSTM with AM***
Layer = 1									
Accuracy (%)	71.67	68.00	71.14	71.71	70.41	69.78	71.63	68.70	70.66
Time	31198.2s	30699.0s	33741.4s	31311.0s	30049.3s	34172.9s	31253.8s	30867.6s	34496.9s
Layer = 2									
Accuracy (%)	71.80	71.37	70.43	71.48	70.52	70.72	71.73	71.02	70.72
Time	34358.2s	33803.2s	39368.4s	34545.9s	30378.9s	39525.9s	35406.5s	34423.0s	51047.1s
Layer = 3									
Accuracy (%)	71.89	70.61	70.49	71.56	69.34	71.14	71.92	71.51	71.07
Time	37518.1s	35790.5s	45315.8s	37431.8s	33869.6s	45603.4s	39672.9s	35405.0s	62172.6s
Layer = 4									
Accuracy (%)	71.76	69.34	70.96	71.45	71.12	70.82	71.74	71.21	71.04
Time	40716.4s	39585.5s	50817.1s	40717.2s	36684.7s	52212.0s	43757.6s	43566.0s	75345.6s
Layer = 5									
Accuracy (%)	71.60	70.47	70.61	71.22	70.63	71.15	71.17	71.56	71.34
Time	43914.6s	39047.6s	55746.6s	44321.7s	39311.7s	57802.3s	61118.7s	59513.1s	87484.8s

Table 5. Accuracy and computational time of PEL with 100 epochs

	Cell Units = 256			Cell Units = 512			Cell Units = 1024		
	SRU with AM	*SRU w/o AM*	*LSTM with AM*	*SRU with AM*	*SRU w/o AM*	*LSTM with AM*	*SRU with AM*	*SRU w/o AM*	*LSTM with AM*
Layer = 1									
Accuracy (%)	85.24	73.85	72.00	87.15	83.68	94.26	95.05	84.77	91.24
Time	33421.9s	32721.2s	36948.1s	27163.3s	24072.0s	29133.1s	31253.8s	30865.0s	35313.6s
Layer = 2									
Accuracy (%)	87.62	72.94	72.25	90.54	74.54	98.12	98.01	85.84	93.38
Time	33449.0s	32912.6s	39997.1s	31546.6s	25420.0s	35143.8s	35406.5s	34550.0s	45639.6s
Layer = 3									
Accuracy (%)	89.21	72.47	74.74	91.21	76.06	98.56	99.92	87.86	99.59
Time	35611.5s	34854.0s	45198.2s	41354.6s	39644.0s	45106.7s	49237.3s	45275.0s	62264.9s
Layer = 4									
Accuracy (%)	87.77	70.65	91.12	90.56	78.29	98.51	99.45	83.93	98.95
Time	41143.1s	39308.6s	46508.4s	43384.1s	42528.0s	50937.3s	53681.9s	51650.0s	75067.7s
Layer = 5									
Accuracy (%)	85.52	71.64	89.00	89.45	84.05	99.47	99.87	84.45	95.99
Time	43986.8s	42539.4s	53273.6s	45413.6s	44088.0s	56948.7s	70870.6s	69115.3s	87870.4s

Table 6. Accuracy and computational time of UTI with 100 epochs

	Cell Units = 256			Cell Units = 512			Cell Units = 1024		
	SRU with AM	*SRU w/o AM*	*LSTM with AM*	*SRU with AM*	*SRU w/o AM*	*LSTM with AM*	*SRU with AM*	*SRU w/o AM*	*LSTM with AM*
Layer = 1									
Accuracy (%)	67.90	64.84	64.38	67.14	60.52	68.19	70.95	64.85	65.81
Time	1753.4s	1720.0s	2034.3s	1845.6s	1809.5s	2056.6s	1922.4s	1833.1s	2115.8s
Layer = 2									
Accuracy (%)	67.62	64.07	66.43	67.62	60.78	64.91	71.62	65.74	69.38
Time	1836.3s	1816.7s	2377.5s	1946.4s	2024.0s	2397.8s	2174.1s	2103.1s	2997.1s
Layer = 3									
Accuracy (%)	68.57	62.84	68.48	68.41	64.81	61.62	72.95	64.19	71.14
Time	2064.6s	2016.2s	2719.8s	2154.3s	2249.5s	2734.9s	2389.0s	2478.6s	3878.4s
Layer = 4									
Accuracy (%)	67.24	63.81	65.24	67.52	64.23	60.57	67.52	68.54	64.38
Time	2369.4s	2224.8s	3058.3s	2436.4s	2453.0s	3049.2s	2642.9s	2673.1s	4547.6s

In Experiment 1, the results are promising and the highest accuracy achieved is 99.99% from PEL while the lowest accuracy is 55.81% from UTI. The overall accuracies from UTI are slightly lower as the dataset is complex due to some similarities in the actions found in different categories. For instance, both punching and hand-shaking involve the movement of a swinging hand, however punching falls under the category of abnormal behavior while hand-shaking is considered as normal behavior. Besides, it is noticeable that higher cell units and number of epochs yield better results. The accuracy is at its peak in the third layer, after stacking more layers like four and five layers into the network, the result shows dropping in accuracy. This indicates that the optimal number of layer for Experiment 1 is three, and more layers results in a reduction of accuracy. Line graph of HCF show as Fig. 7 and Line graph of PEL show as Fig. 8.

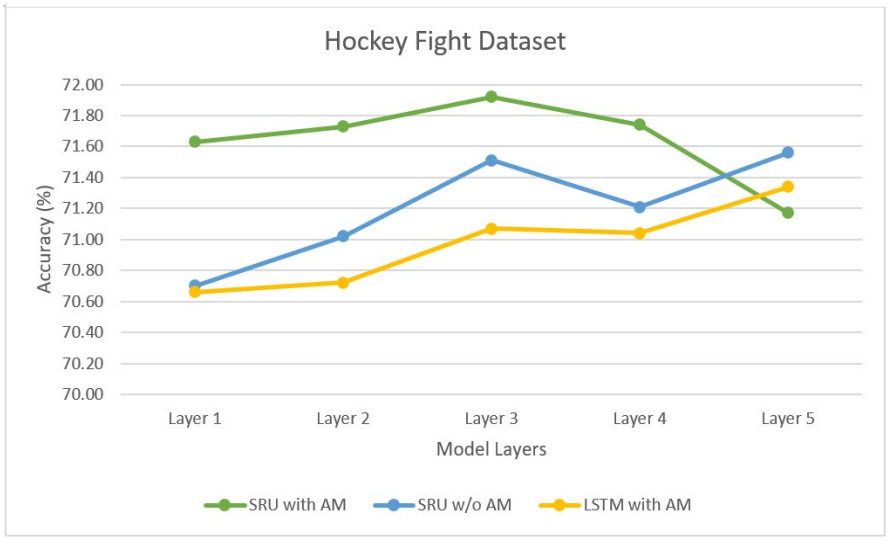

Fig. 7. Line graph of HCF with accuracy according to each layer with cell units of 1024

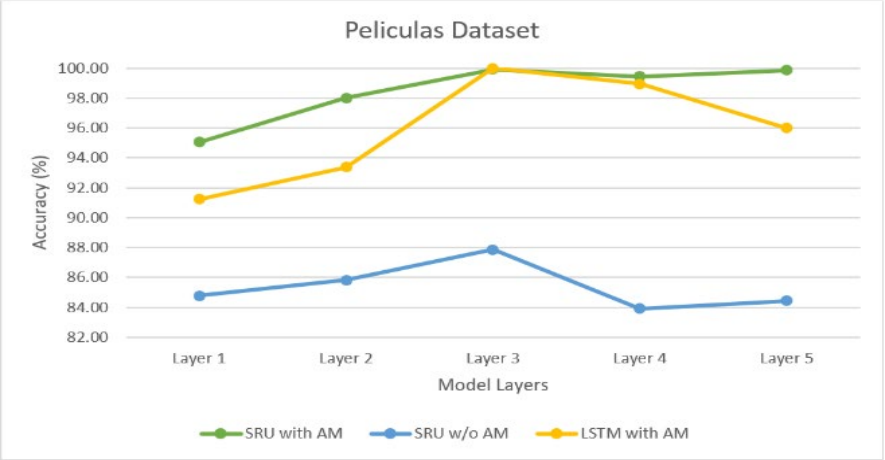

Fig. 8. Line graph of PEL with accuracy according to each layer with cell units of 1024

In addition, it is observed that the performance of SRU with AM is better compared to SRU without AM and LSTM with AM in terms of accuracy although the time taken for training is shorter without AM. From Fig. 9 - Fig. 10, the computational time taken has been vastly reduced by 1.5 times using SRU compared to LSTM. Other than that, it is noticeable that by staking more layers of SRU, the time consumption of model training in comparison with LSTM is significantly reduced by around 67%.

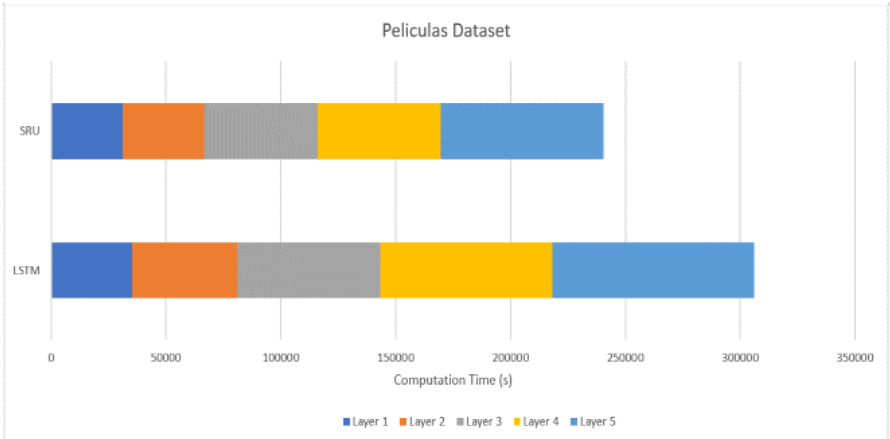

Fig. 9. Bar charts of PEL with computation time according to number of layers

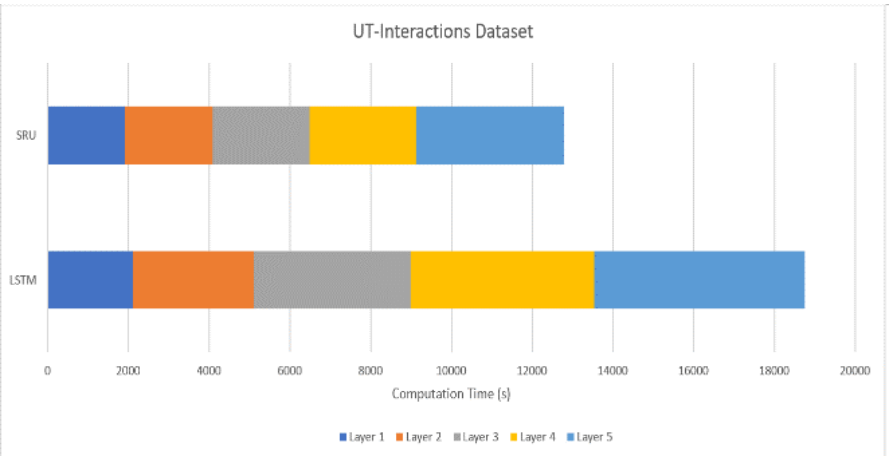

Fig. 10. Bar charts of UTI with computation time according to number of layers

4.2. Experiment 2

Experiment 2 involves crowd-based interactions such as fleeing from group or crowd fighting. This study looks into the robustness of model in identifying abnormal behaviours among a larger group of people. Attention mechanism is applied during the stage of feature extraction to determine where are the focus of the model when extracting the salient features from the images. Fig. 11 and Fig. 12 represent some of the selected video frames from Experiment 2 whereby the upper row indicates the feature map of SRU without attention mechanism and the lower row represents the feature map of SRU with attention mechanism.

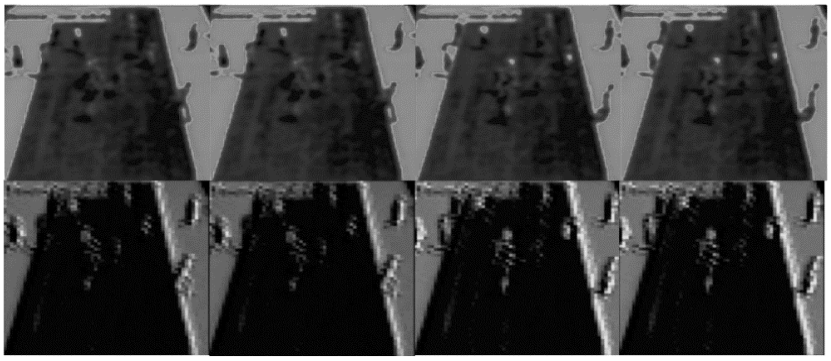

Fig. 11. Visualization of UMN dataset

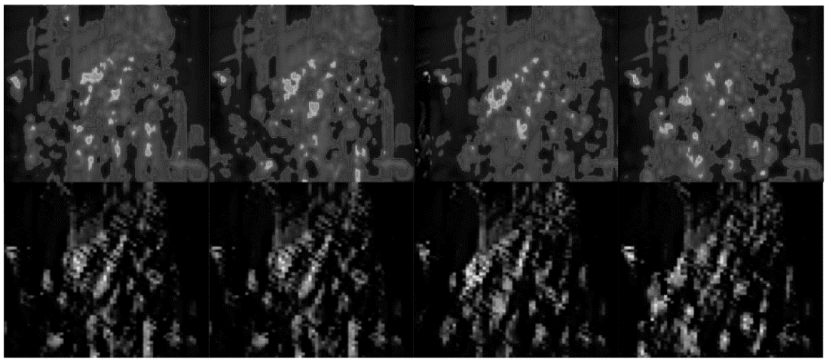

Fig. 12. Visualization of WEB dataset

In Fig. 11, it is seen that without attention mechanism, the feature map is ambiguous whereas, with attention mechanism, it focuses on the crowd that are fleeing from the setting. In Fig. 12, the focus of the model is sparse as compared to the one with attention mechanism. SRU with attention mechanism is able to scope down the area and focus only on salient part of the images which are the human subjects

that are running on a street. The features are then passed to the classifiers to determine the temporal information of the videos. Experiment 2 achieves significant results with 99.87% being the highest as there is significant difference between the temporal information of abnormal behavior and normal behavior. In UMN and WEB datasets, both involve crowd fleeing scenes as abnormal behavior and crowd walking as normal behavior. It is observed from frame to frame that the movement of the crowd fleeing is very chaotic and obvious compared to the less noticeable pedestrians walking with slow movement and pace. The accuracies of the model for various number of epochs, cell units and layers based on 0.001 learning rate are summarized in Table 7 and Table 8, Fig. 13 and Fig. 14 for SRU with and without AM, and LSTM with AM.

Table 7. Accuracy and computational time of UMN with 100 epochs

	Cell Units = 256			Cell Units = 512			Cell Units = 1024		
	SRU with AM	*SRU w/o AM*	*LSTM with AM*	*SRU with AM*	*SRU w/o AM*	*LSTM with AM*	*SRU with AM*	*SRU w/o AM*	*LSTM with AM*
Layer = 1									
Accuracy (%)	96.28	94.26	96.41	97.08	95.08	97.03	98.13	94.53	98.09
Time	31495.6s	30318.2s	33478.2s	31456.0s	30685.6s	34156.9s	31253.8s	29232.6s	34156.2s
Layer = 2									
Accuracy (%)	97.20	94.52	96.49	97.12	94.06	97.10	98.34	94.29	98.18
Time	34596.1s	33788.2s	39547.6s	34562.0s	33831.6s	39586.4s	35478.1s	33783.2s	51478.6s
Layer = 3									
Accuracy (%)	98.25	94.81	98.16	98.69	94.81	98.65	98.76	94.32	98.33
Time	37451.2s	34258.2s	45123.0s	37489.6s	37316.4s	45623.1s	39548.2s	38889.4s	62153.4s
Layer = 4									
Accuracy (%)	97.65	94.91	97.40	98.21	95.08	98.56	98.71	94.81	98.03
Time	40125.9s	36415.4s	50147.2s	40157.9s	40752.8s	522146.3	43156.2s	42301.6	75489.6s
Layer = 5									
Accuracy (%)	97.76	94.34	96.72	98.21	93.16	97.76	98.70	94.25	98.00
Time	43269.1s	40047.8s	55478.2s	44156.2s	44624.8s	57489.2s	61472.6s	59356.8s	87153.3s

Table 8. Accuracy and computational time of WEB with 100 epochs

	Cell Units = 256			Cell Units = 512			Cell Units = 1024		
	SRU with AM	*SRU w/o AM*	*LSTM with AM*	*SRU with AM*	*SRU w/o AM*	*LSTM with AM*	*SRU with AM*	*SRU w/o AM*	*LSTM with AM*
Layer = 1									
Accuracy (%)	87.75	88.33	83.12	91.00	90.90	84.92	86.12	89.40	87.92
Time	32800.9s	32005.6s	34829.0s	39153.1s	36410.1s	42265.6s	39321.5s	37456.6s	41942.3s
Layer = 2									
Accuracy (%)	89.32	85.58	83.76	91.12	86.79	86.69	88.28	89.29	89.39
Time	40665.5s	35840.5s	45142.9s	40459.0s	41413.4s	45236.1s	50313.1s	49211.9s	54984.7s
Layer = 3									
Accuracy (%)	98.32	86.55	84.76	95.40	85.12	91.92	99.87	92.98	93.80
Time	44684.6s	41541.2s	50043.6s	53214.8s	44323.3s	59156.3s	55123.6s	52712.2s	61138.5s
Layer = 4									
Accuracy (%)	93.98	86.15	87.77	94.75	85.36	92.92	92.54	92.33	93.93
Time	51669.6s	45661.8s	60344.3s	60156.5s	56456.7s	71626.6s	64734.4s	60156.3s	74569.1s
Layer = 5									
Accuracy (%)	91.35	85.39	88.36	92.48	88.36	94.72	91.40	92.35	93.98
Time	64831.5s	63463.0s	76235.8s	66045.2s	58161.2s	78166.3s	72316.0s	70145	84654.6s

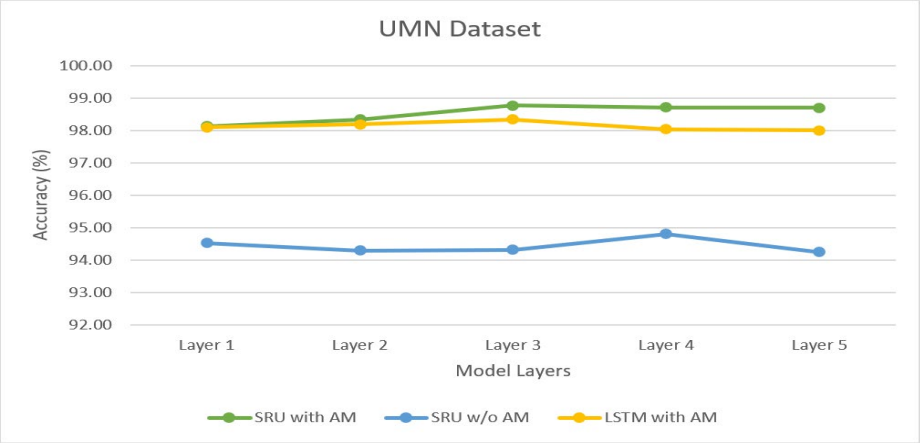

Fig. 13. Line graphs of UMN with accuracy according to each layer with cell units of 1024

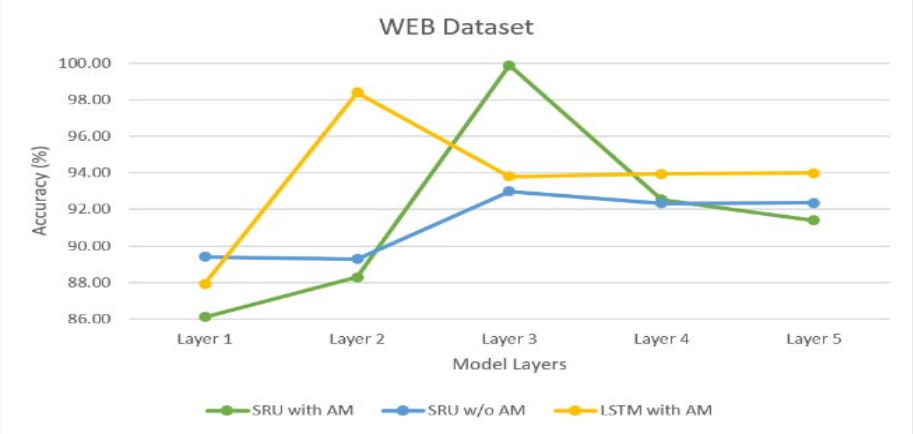

Fig. 14. Line graphs of WEB with accuracy according to each layer with cell units of 1024

The results from Experiment 2 show encouraging performance with the highest accuracy of 99.87% from WEB. The overall performance of UMN dataset is high with an average accuracy of around 90% and above. This is due to the fact that the videos are uniform in terms of the type of settings and movement patterns. Similar to Experiment 1, the accuracy is at its peak in the third layer. Other than that, it is observed that the performance of SRU is slightly better as compared to LSTM with a speed up of 1.5 times. The computational efficiency of SRU is clearly increased when more layers are stacked to it. The bar chart of UMN is shown in Fig. 15 and the Bar chart of WEB is shown in Fig. 16.

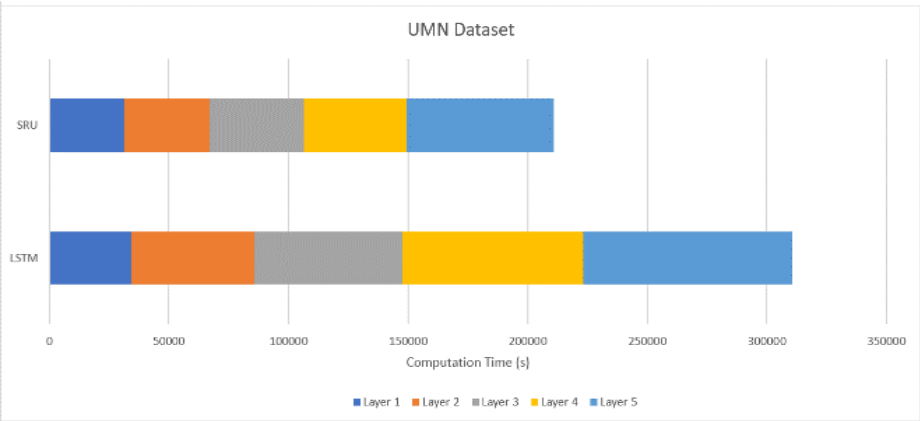

Fig. 15. Bar chart of UMN with computation time according to number of layers

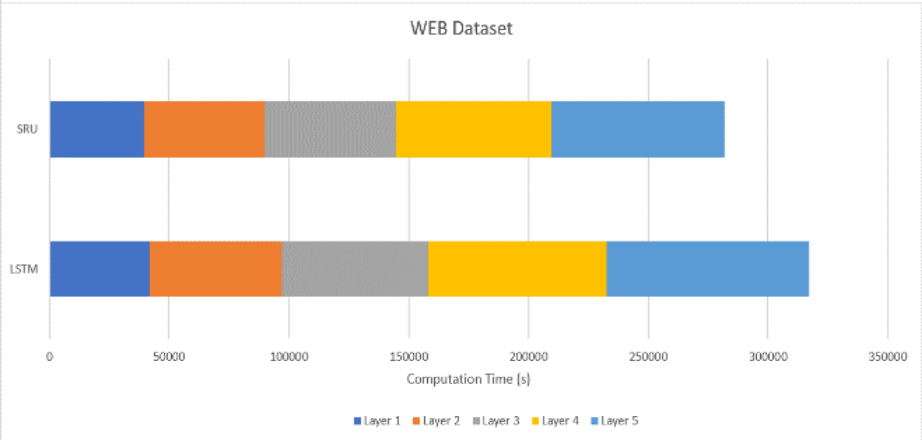

Fig. 16. Bar chart of WEB with computation time according to number of layers

The development and validation of our model mark a significant advancement in the field of intelligent surveillance systems. The empirical results demonstrate not only the model's superior performance in accurately identifying abnormal behaviors across diverse scenarios but also its computational efficiency, a crucial factor for real-time application in public safety infrastructure. Our findings suggest that the integration of an SRU with an attention mechanism can effectively address the limitations of traditional LSTM-based models, particularly in terms of processing speed and the ability to focus on relevant aspects of the video data. This has profound implications for the deployment of intelligent surveillance systems in real-world settings, where the timely and accurate detection of abnormal behaviors can significantly enhance public safety measures.

5. Conclusion

In this study, we conducted two sets of experiments to evaluate the effectiveness of the attention mechanism and SRU in detecting abnormal behaviors. In scenarios involving interactions between two individuals, the attention mechanism predominantly concentrates on human body lines to decipher actions. Conversely, in crowd-based interaction scenarios, the focus shifts more broadly to the human figures, although the delineation of body lines is less distinct due to the proximity of individuals in crowded settings. Nonetheless, the analysis extends beyond singular images, examining body movement patterns and attention focus across sequential frames for prediction. When contrasted with LSTM models, our proposed method demonstrated superior accuracy and computational efficiency. Notably, layering additional SRU units reduced training time by a factor of 1.5 compared to LSTM, confirming SRU's aptness for rapid and accurate abnormal behavior detection. Regarding soft attention's static glimpse position, future explorations will consider alternative attention mechanisms, like the spatial transformer and hard attention, to investigate their impact on model visualization and glimpse positioning adjustments. This research applied the proposed methods to offline video sequences, positioning them primarily within forensic and investigative contexts. Future work will aim to broaden the application to real-time video analysis, with a focus on developing an instantaneous notification system for immediate alerts upon detection of abnormal behaviors.

Acknowledgment

The authors thank the Ministry of Higher Education, Malaysia for supporting the research through the Fundamental Research Grant Scheme (FRGS) with grant number: FRGS/1/2020/ICT02/MMU/02/5.

Declarations

Author contribution. All authors contributed equally to the main contributor to this paper. All authors read and approved the final paper.

Funding statement. This work was supported by the Fundamental Research Grant Scheme (FRGS) under the Ministry of Higher Education, Malaysia (Grant number: FRGS/1/2020/ICT02/MMU/02/5).
Conflict of interest. The authors declare no conflict of interest.
Additional information. No additional information is available for this paper.

References

[1] S. Roka, M. Diwakar, P. Singh, and P. Singh, "Anomaly behavior detection analysis in video surveillance: a critical review," *J. Electron. Imaging*, vol. 32, no. 04, p. 042106, Mar. 2023, doi: 10.1117/1.JEI.32.4.042106.

[2] M. Cho, T. Kim, W. J. Kim, S. Cho, and S. Lee, "Unsupervised video anomaly detection via normalizing flows with implicit latent features," *Pattern Recognit.*, vol. 129, p. 108703, Sep. 2022, doi: 10.1016/j.patcog.2022.108703.

[3] L. Wang, H. Tan, F. Zhou, W. Zuo, and P. Sun, "Unsupervised Anomaly Video Detection via a Double-Flow ConvLSTM Variational Autoencoder," *IEEE Access*, vol. 10, pp. 44278–44289, 2022, doi: 10.1109/ACCESS.2022.3165977.

[4] X. Wang *et al.*, "Robust Unsupervised Video Anomaly Detection by Multipath Frame Prediction," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 33, no. 6, pp. 2301–2312, Jun. 2022, doi: 10.1109/TNNLS.2021.3083152.

[5] C. Huang *et al.*, "Self-Supervised Attentive Generative Adversarial Networks for Video Anomaly Detection," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 34, no. 11, pp. 9389–9403, Nov. 2023, doi: 10.1109/TNNLS.2022.3159538.

[6] R. Leider, "The Modern Common Law of Crime," *J. Crim. Law Criminol.*, vol. 111, no. 2, pp. 407–499, Jan. 2021. [Online]. Available at: https://scholarlycommons.law.northwestern.edu/jclc/vol111/iss2/2.

[7] T. Abam, "Impact of Terrorism on Society Insecurities," *J. Anthropol. Reports*, vol. 5, no. 5, pp. 9–10, Sep. 2022. [Online]. Available at: https://www.walshmedicalmedia.com/open-access/impact-of-terrorism-on-society-insecurities-114620.html.

[8] A. Birze, K. Regehr, and C. Regehr, "Workplace Trauma in a Digital Age: The Impact of Video Evidence of Violent Crime on Criminal Justice Professionals," *J. Interpers. Violence*, vol. 38, no. 1–2, pp. 1654–1689, Jan. 2023, doi: 10.1177/08862605221090571.

[9] S. Kim, P. Joshi, P. S. Kalsi, and P. Taheri, "Crime Analysis Through Machine Learning," in *2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, Nov. 2018, pp. 415–420, doi: 10.1109/IEMCON.2018.8614828.

[10] N. Carmack, "Benefits of Surveillance Cameras in Public Places," *BOS Security*, 2022. [Online]. Available at: https://www.bossecurity.com/2022/12/21/benefits-of-surveillance-cameras-in-public-places/.

[11] I. Insider, "Role of CCTV Cameras: Public, Privacy and Protection," *IFSEC Insider | Security and Fire News and Resources*, 2021. [Online]. Available at: https://www.ifsecglobal.com/video-surveillance/role-cctv-cameras-public-privacy-protection/.

[12] W. Sultani, C. Chen, and M. Shah, "Real-World Anomaly Detection in Surveillance Videos," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2018, pp. 6479–6488, doi: 10.1109/CVPR.2018.00678.

[13] T. Zhang *et al.*, "Recent Advances in Video Analytics for Rail Network Surveillance for Security, Trespass and Suicide Prevention—A Survey," *Sensors*, vol. 22, no. 12, p. 4324, Jun. 2022, doi: 10.3390/s22124324.

[14] R. Nawaratne, D. Alahakoon, D. De Silva, and X. Yu, "Spatiotemporal Anomaly Detection Using Deep Learning for Real-Time Video Surveillance," *IEEE Trans. Ind. Informatics*, vol. 16, no. 1, pp. 393–402, Jan. 2020, doi: 10.1109/TII.2019.2938527.

[15] B. Kiran, D. Thomas, and R. Parakkal, "An Overview of Deep Learning Based Methods for Unsupervised and Semi-Supervised Anomaly Detection in Videos," *J. Imaging*, vol. 4, no. 2, p. 36, Feb. 2018, doi: 10.3390/jimaging4020036.

[16] T. Alanazi, K. Babutain, and G. Muhammad, "A Robust and Automated Vision-Based Human Fall Detection System Using 3D Multi-Stream CNNs with an Image Fusion Technique," *Appl. Sci.*, vol. 13, no. 12, p. 6916, Jun. 2023, doi: 10.3390/app13126916.

[17] A. Lentzas and D. Vrakas, "Non-intrusive human activity recognition and abnormal behavior detection on elderly people: a review," *Artif. Intell. Rev.*, vol. 53, no. 3, pp. 1975–2021, Mar. 2020, doi: 10.1007/s10462-019-09724-5.

[18] Honghai Liu, Shengyong Chen, and N. Kubota, "Intelligent Video Systems and Analytics: A Survey," *IEEE Trans. Ind. Informatics*, vol. 9, no. 3, pp. 1222–1233, Aug. 2013, doi: 10.1109/TII.2013.2255616.

[19] C. Liu, Y. Zhang, Y. Xue, and X. Qian, "AJENet: Adaptive Joints Enhancement Network for Abnormal Behavior Detection in Office Scenario," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 3, pp. 1427–1440, Mar. 2024, doi: 10.1109/TCSVT.2023.3295432.

[20] H.-T. Duong, V.-T. Le, and V. T. Hoang, "Deep Learning-Based Anomaly Detection in Video Surveillance: A Survey," *Sensors*, vol. 23, no. 11, p. 5024, May 2023, doi: 10.3390/s23115024.

[21] N. C. Tay, T. Connie, T. S. Ong, A. B. J. Teoh, and P. S. Teh, "A Review of Abnormal Behavior Detection in Activities of Daily Living," *IEEE Access*, vol. 11, pp. 5069–5088, 2023, doi: 10.1109/ACCESS.2023.3234974.

[22] S.-H. Cho and H.-B. Kang, "Abnormal behavior detection using hybrid agents in crowded scenes," *Pattern Recognit. Lett.*, vol. 44, pp. 64–70, Jul. 2014, doi: 10.1016/j.patrec.2013.11.017.

[23] D. Bahdanau, K. H. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc.*, pp. 1–15, 2015, [Online]. Available at: https://arxiv.org/abs/1409.0473.

[24] E. Bermejo Nievas, O. Deniz Suarez, G. Bueno García, and R. Sukthankar, "Violence Detection in Video Using Computer Vision Techniques," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 6855 LNCS, no. PART 2, Springer, Berlin, Heidelberg, 2011, pp. 332–339. [Online]. Available at: 10.1007/978-3-642-23678-5_39.

[25] Nievas *et al.*, "Computer Analysis of Images and Patterns," *Academic Torrents*, pp. 332-339. 2011. [Online]. Available at: https://academictorrents.com/details/70e0794e2292fc051a13f05ea6f5b6c16f3d3635.

[26] Ryoo, "UT-Interaction Dataset," *Papers With Code*, 2020. [Online]. Available at: https://paperswithcode.com/dataset/ut-interaction.

[27] D. N. Papanikolopoulos *et al.*, "Monitoring Human Activity," *University of Minnesota* [Online]. Available at: https://mha.cs.umn.edu/.

[28] K. Soomro, "UCF101 - Action Recognition Data Set," *Center for Research in Computer Vision at the University of Central Florida*, 2013. https://www.crcv.ucf.edu/data/UCF101.php.

[29] N. C. Tay, C. Tee, T. S. Ong, and P. S. Teh, "Abnormal Behavior Recognition using CNN-LSTM with Attention Mechanism," in *2019 1st International Conference on Electrical, Control and Instrumentation Engineering (ICECIE)*, Nov. 2019, pp. 1–5, doi: 10.1109/ICECIE47765.2019.8974824.

[30] T. Lei, Y. Zhang, S. I. Wang, H. Dai, and Y. Artzi, "Simple Recurrent Units for Highly Parallelizable Recurrence," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 4470–4481, doi: 10.18653/v1/D18-1477.

[31] K.-E. Ko and K.-B. Sim, "Deep convolutional framework for abnormal behavior detection in a smart surveillance system," *Eng. Appl. Artif. Intell.*, vol. 67, pp. 226–234, Jan. 2018, doi: 10.1016/j.engappai.2017.10.001.

[32] D. Arifoglu and A. Bouchachia, "Activity Recognition and Abnormal Behaviour Detection with Recurrent Neural Networks," *Procedia Comput. Sci.*, vol. 110, pp. 86–93, Jan. 2017, doi: 10.1016/j.procs.2017.06.121.

[33] Y. Fan, G. Wen, D. Li, S. Qiu, and M. D. Levine, "Early event detection based on dynamic images of surveillance videos," *J. Vis. Commun. Image Represent.*, vol. 51, pp. 70–75, Feb. 2018, doi: 10.1016/j.jvcir.2018.01.002.

[34] S. Sharma, R. Kiros, and R. Salakhutdinov, "Action Recognition using Visual Attention," in *ICLR 2016 - 10th International Conference on Learning Representations*, 2016, pp. 1–6, [Online]. Available at: http://arxiv.org/abs/1511.04119.

[35] X. Chen, J. Yu, and Z. Wu, "Temporally Identity-Aware SSD With Attentional LSTM," *IEEE Trans. Cybern.*, vol. 50, no. 6, pp. 2674–2686, Jun. 2020, doi: 10.1109/TCYB.2019.2894261.

[36] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-Scale Video Classification with Convolutional Neural Networks," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2014, pp. 1725–1732, doi: 10.1109/CVPR.2014.223.

[37] L. Zhao, L. Zhu, S. Zhao, and X. Ma, "Sequestration and bioavailability of perfluoroalkyl acids (PFAAs) in soils: Implications for their underestimated risk," *Sci. Total Environ.*, vol. 572, pp. 169–176, Dec. 2016, doi: 10.1016/j.scitotenv.2016.07.196.

[38] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, "Spatial Transformer Networks," *Adv. Neural Inf. Process. Syst.*, vol. 2015-January, pp. 2017–2025, Jun. 2015. [Online]. Available at: https://arxiv.org/abs/1506.02025v3.

[39] G. F. Elsayed, S. Kornblith, and Q. V. Le, "Saccader: Improving accuracy of hard attention models for vision," *Adv. Neural Inf. Process. Syst.*, vol. 32, no. NeurIPS, p. 13, 2019, [Online]. Available at: https://arxiv.org/abs/1908.07644.

[40] David, "Not afraid of 'overfitting,'" 2024. [Online]. Available at: http://nooverfit.com/wp/.

[41] K. Greff, R. K. Srivastava, J. Koutnik, B. R. Steunebrink, and J. Schmidhuber, "LSTM: A Search Space Odyssey," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 28, no. 10, pp. 2222–2232, Oct. 2017, doi: 10.1109/TNNLS.2016.2582924.

[42] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "HMDB: A large video database for human motion recognition," in *2011 International Conference on Computer Vision*, Nov. 2011, pp. 2556–2563, doi: 10.1109/ICCV.2011.6126543.