




# CNN features with bi-directional LSTM for real-time anomaly detection in surveillance networks

Waseem Ullah<sup>1</sup> · Amin Ullah<sup>1</sup> · Ijaz Ul Haq<sup>1</sup> · Khan Muhammad<sup>2</sup> · Muhammad Sajjad<sup>3</sup> · Sung Wook Baik<sup>1</sup> 

Received: 15 January 2020 / Revised: 14 July 2020 / Accepted: 21 July 2020 /

Published online: 20 August 2020

© Springer Science+Business Media, LLC, part of Springer Nature 2020

## Abstract

In current technological era, surveillance systems generate an enormous volume of video data on a daily basis, making its analysis a difficult task for computer vision experts. Manually searching for unusual events in these massive video streams is a challenging task, since they occur inconsistently and with low probability in real-world surveillance. In contrast, deep learning-based anomaly detection reduces human labour and its decision making ability is comparatively reliable, thereby ensuring public safety. In this paper, we present an efficient deep features-based intelligent anomaly detection framework that can operate in surveillance networks with reduced time complexity. In the proposed framework, we first extract spatiotemporal features from a series of frames by passing each one to a pre-trained Convolutional Neural Network (CNN) model. The features extracted from the sequence of frames are valuable in capturing anomalous events. We then pass the extracted deep features to multi-layer Bi-directional Long Short-term Memory (BD-LSTM) model, which can accurately classify ongoing anomalous/normal events in complex surveillance scenes of smart cities. We performed extensive experiments on various anomaly detection benchmark datasets to validate the functionality of the proposed framework within complex surveillance scenarios. We reported a 3.41% and 8.09% increase in accuracy on UCF-Crime and UCFCrime2Local datasets compared to state-of-the-art methods.

**Keywords** Anomaly detection · Deep learning · LSTM · Intelligent surveillance networks · Smart surveillance · Crime detection

---

✉ Sung Wook Baik  
sbaik@sejong.ac.kr

<sup>1</sup> Intelligent Media Laboratory, Digital Contents Research Institute, Sejong University, Seoul, Republic of Korea

<sup>2</sup> Department of Software, Sejong University, Seoul, Republic of Korea

<sup>3</sup> Department of Computer Science, Islamia College Peshawar, Peshawar, Pakistan

## 1 Introduction

In real-world surveillance scenarios, the anomaly detection such as fighting, stealing and robbery etc. is attracting computer vision researchers due to its daily life surveillance applications. Recently, a huge amount of surveillance cameras are deployed worldwide in variable locations for public safety. These cameras continuously generate huge amount of video data, requiring monitoring efforts that are mostly performed by humans, which is tedious and erroneous, thereby creating the need of automatic monitoring techniques. Due to limited performance of manual monitoring, law enforcement agencies are showing miserable performance in to capturing or avoiding abnormal events. To detect anomalous activities, an efficient computer vision-based method is required that can classify normal/abnormal events effectively without human interference. Such an automatic method is not only helpful in monitoring, but it also reduces the human effort required to maintain manual observation on a 24-h basis. There are several methods in the literature [1, 5, 9, 34, 40] in which anomalous activities are defined as “the occurrence of variation in normal patterns”. Using this definition, anomalous event detection is studied as a classification problem [6, 7, 35] in which visual information is fed into the classifier to make it learn about the difference of normal and abnormal events. In some cases, it is considered as a binary classification problem, for example violence detection [16, 37] and traffic road accident detection [2, 18]. However, these techniques are limited to only two types of activities: violent/non-violent or accident/normal and hence providing a fractional solution for practice in real-world scenarios.

The performance of these techniques is relatively lower in real-world scenarios, since the diverse and dense nature of surveillance data makes it very challenging to detect all possible anomalous events. Most of the existing methods depend on information about previous events, and it is therefore desirable that new anomaly detection systems do not depend upon such information. Previously, researchers have used either traditional features with machine learning techniques or deep learning techniques for anomaly detection [6, 20, 29, 38]. Anomaly detection using sparse coding techniques [7, 13, 20] achieved good results so far, and such approaches are considered to be representative techniques for anomaly detection. These models are trained in such a way that the initial parts of the video clips (i.e., before the occurrence of an anomalous event) are used to create a dictionary for normal events. However, this is an inadequate method to allow the system to accurately detect anomalous events from a dictionary trained on normal events. The multi-instance learning (MIL) method based on weakly supervised techniques is also used for anomaly detection [13, 14, 31]. In this approach, the videos are divided into a predetermined number of clips in the training stage. These clips create instances of bags, consisting of both positive (anomalous event) and negative (normal events) bags, and can learn instance-level labels.

Since surveillance environments change over time (e.g., at various times of day), techniques based on sparse coding have certain limitations, for instance, changing the dictionary learned from normal events to anomalous events results in a high false alarm rate. Furthermore, detecting anomalous events in surveillance videos is extremely difficult due to their low resolution, large intra/inter class adaptation and lack of annotated data, since these occurrences come off rarely related to normal appearance. Humans can recognise typical or unusual events based on common sense, but machines need to use visual features to detect these events. In general, stronger visual features perform better in terms of the detection and recognition of events [41].

Mostly, the existing techniques are suffering from the problem of high false alarms rate. Furthermore, these techniques work well on simple datasets, however, their performance is limited when they deal with the real-life scenarios. To handle these issues, in this paper we develop a robust and efficient model which learns visual features from a sequence of 15 frames by integrating them in the form of spatiotemporal information from raw video. We utilise a weakly supervised method based on spatiotemporal features and BD-LSTM to train our model. The key contributions of current work can be summarised as follows:

- We propose an efficient CNN-based intelligent paradigm for anomaly detection functional in both indoor and outdoor surveillance networks. We utilise a pre-trained ResNet-50 architecture for deep spatiotemporal features extraction followed by a sequential learning method that outperforms state-of-the-art approaches in terms of accuracy.
- A multi-layer BD-LSTM architecture is utilised in the proposed framework. This boosts the capabilities of effective learning due to the forward and backward pass applied at each layer of the LSTM model. The final trained model is therefore not limited to the data used for training, but it is also functional in video surveillance networks. We therefore contribute to the existing anomaly detection literature by presenting a framework with a high level of adaptability to real-world smart city surveillance scenarios.
- Our proposed framework is evaluated using various challenging benchmark datasets. Unlike existing anomaly detection methods, we accomplish state-of-the-art outcomes by using 2D CNNs with reduced model size and fewer parameters and having the size of 143 MBs; this scheme allows real-time and precise anomaly detection, as it takes only 0.20s to process a single 15 frames sequence.

The remainder of the paper is organized as follows. Section 2 discusses review of existing techniques in the literature. Section 3 presents the explanation of the overall proposed framework. Section 4 evaluates the experimental results of our research and provides the comparison with existing techniques, followed by a conclusion and suggestions for future research directions in Section 5.

## 2 Related work

Anomalous events detection in surveillance videos has been studied widely for several years and is a challenging problem due to visual features' variations and inter/intra class differences. Numerous researchers have addressed this problem based on the hypothesis that abnormalities are unseen or infrequent, and that changes in normal behaviour are assumed to be unusual patterns. The literature on anomaly detection can be mostly classified into three categories: statistical features [6, 15, 16, 18]; deep features [20, 21, 27, 39]; and ranking based methods [10, 31, 41]. In the following sections, the literature is discussed in detail.

### 2.1 Statistical features-based methods

Numerous statistical features-based models are recommended for anomaly detection in early literature within video footage, such as Gaussian process models [6, 18] and hidden Markov

models [8, 16]. Mehran et al. [24] proposed a social force paradigm in which the collaboration forces were calculated, applying optical flow to detect normal as well as anomalous events. Similarly, Kim and Grauman [15] suggested a system based on a Markov Random Field and used the spatiotemporal domain to detect local and global anomalies. Li et al. [18] proposed a detector that used a mixture of dynamic texture models for anomaly detection in crowds. Sparse reconstruction approaches are utilized anomalies detection and outliers in pattern modelling [21]. Another approach, Cong et al. [9] proposed the concept of Sparse Reconstruction Cost (SRC) for the detection of abnormal events, which used sequence of images or patches of local spatiotemporal features to spot both local and global unusual occurrences. Another unsupervised approach [39] utilised fully dynamic sparse coding style to distinguish abnormalities in videos based on online query signals and sparse reconstruct ability obtained from a learned dictionary of all events. However, learning to detect abnormalities in a timely manner has remained challenging, and this problem has attracted the attention of many researchers. For instance, Lu et al. [20] applied a Sparse Combination Learning (SCL) framework using an efficient sparse approach and analysed their method by using both local and cloud servers.

## 2.2 Deep learning-based methods

In recent times, deep learning has reached human levels of accuracy in several computer vision applications including surveillance [4, 25]. Similarly, many researchers have utilized deep features to detect anomalous events in videos. For instance, the authors of [20, 39] utilised deep autoencoders to understand typical behaviours and used reconstruction loss to detect anomalies. Luo et al. [21] introduced a technique for temporal coherence-based abnormality revealing based on sparse coding, including processing carried out by a stacked Recurrent Neural Network (RNN). Sabokrou et al. [27] integrated CNNs with 3D deep autoencoders in order to detect anomalies in videos. Some researchers have designed deep neural networks for abstraction and feature learning [8, 11, 22] and video prediction learning for anomalies [19]. Hasan et al. [11] presented two methods for Learning Temporal Regularity (LTR) in video sequences using autoencoders. They first adopted traditional handcrafted features for a sequence of video frames and trained them on a fully connected autoencoder, and then presented an end-to-end feed-forward fully convolutional autoencoder to learn regularity. Finally, they evaluated their method in both qualitative and quantitative ways. In another study, Chong et al. [8] recommended a novel approach for abnormal event identification via spatiotemporal autoencoders. The main aim of their method was to learn spatiotemporal features using an efficient autoencoder and to allow it to process up to 140 frames per second. Luo et al. [22] applied a CNN and ConvLSTM for anomaly detection based on the concept of memorising all past frames resembling motion features. They also combined an autoencoder with ConvLSTM to encode the motion and appearance of objects inside diverse events. Another method suggested by Huo et al. [14] for the identification of abnormal events in videos exploited multi-instance dictionary learning. Their system is appropriate in cases where the label for a set of sub-events is easily obtainable, while the sub-event labels are uncertain. Sultani et al. [31] introduced a method for learning anomalies via a deep multi-instance ranking technique by way of leveraging indecisively labelled training videos, i.e. with training labels at video level rather than clip level.

## 2.3 Ranking-based methods

Over the past few years, rank-based anomaly detection methods have attracted many researchers due to their effective performance. These techniques have mostly focused on increasing the relative scores of elements rather than the individual scores. Weakly supervised MIL methods are utilized for anomaly detection [31]. For instance, Le et al. [13] recommended an anomalous events detection paradigm based on learning rank for abnormal event detection by utilising anomaly information, and a MIL scheme was developed based on a graph. After discovering positive occurrences, they trained a kernel support vector machine as a course filter and improved dictionary learning for anomaly detection.

Sultani et al. [31] introduced a MIL-based anomaly detection, and trained their model on both normal and anomalous events. They used deep anomaly ranking for video-level labelling to predict anomaly scores. In this approach, two different bags were used to separate normal and anomalous videos, and a MIL was used to predict anomaly scores for the videos. Landi et al. [17] trained a regression model for anomalies, using a tube extraction method and set coordinates. The input tube was applied to determine the action representation in the input video and hybrid features were then extracted, such as features from inception block and optical flow, which are further integrated using an average pooling layer and then fed into the regression network. Zhong et al. [40] introduced a model-based weakly supervised anomaly detection method and a fully supervised model for action classification with noisy labels. In this approach, only the labels for anomaly videos were noisy due to unknown anomalous events. To clean these noisy labels, a graph CNN was trained and an action classifier was used to classify the activities. Table 1 presents a summary of the recent literature in terms of the main contributions of each study, the domain of application and the datasets used.

## 3 Proposed methodology

The overall proposed framework and its key elements are deliberated in this section based on the data flow demonstrated in Fig. 1. This framework is divided into three main steps: a single

**Table 1** Anomaly detection techniques for surveillance monitoring with a summary of their main contributions

Paper	Techniques	Main contributions	Datasets
[31] (2018)	Multi-instance ranking scores for anomaly prediction in surveillance video clips	The MIL ranking method uses bags of videos to understand a deep anomaly ranking paradigm that calculates anomalous events based on high anomaly scores	UCF-Crime [31]
[13] 2018	MIL-based technique for normal and abnormal event detection	Graph-based MIL detection approaches for both abnormal and normal data in videos	UCSD pedestrian [23]
[40] 2019	Supervised learning with noisy labels	Two supervised learning approaches: (i) cleaning noisy labels; (ii) detecting weak anomalies by employing an action classifier	UCF-Crime [31] UCSD pedestrian [23] ShanghaiTech Campus [19]
[17] (2019)	Action-based tube for anomaly localisation	Localises anomalies in the video using a spatiotemporal tube	UCFCrime2Local [17]
[41] (2019)	Attention-based MIL temporal augment network	MIL ranking model with motion features using a temporal augmented network.	UCF-Crime [31]

frame from the surveillance video is fed into a pre-trained ResNet-50 model to extract features; a feature vector is generated from a 15 consecutive frames of the video; the acquired feature vector is passed to a multi-layer BD-LSTM to recognise anomalous events. Each step is discussed in the subsequent section in detail.

### 3.1 Feature extraction using a pre-trained ResNet-50

A deep learning model needs a huge number of images to train it from scratch, and also requires high-capacity processing units. To overcome this problem, researchers have utilised transfer learning techniques to obtain a suitable model for a specific task, and pre-trained model’s weights are utilised to fine-tune a model for other applications than the original one [12]. In the proposed framework, we also use pre-trained ResNet-50 residual networks (ResNets) for feature extraction, which are trained on the ImageNet dataset. ResNet-50 is a deep CNN model in which the essential concept is to skip one or more layers by utilizing shortcut connections. The fundamental cell blocks in this network are called “bottlenecks” and follow these rules: the similar number of filters in a layer have the equivalent number of output feature maps, and if the size of the feature map is reduced, the amount of filters is doubled up. Down-sampling is achieved through a convolutional layer with stride two, followed by batch normalisation prior to application of the ReLU activation function. When the dimensions of the input and output are different, shortcut projection is applied to fit up the dimensions with the  $1 \times 1$  convolutions, while an identity shortcut is employed if the dimensions are the same [12, 36]. ResNet-50 has a total of 50 weighted number of layers, with 23.5 million trainable parameters. We extracted features from the last fully connected layer of a 15-frames sequence, which are then fed into the BD-LSTM for further processing. The basic structural design of the ResNet-50 is demonstrated in Fig. 2.

### 3.2 Recurrent neural networks

RNNs are renowned for their competence towards explore hidden sequential information in equally spatial and temporal sequential data. A video consists of a series of frames that provides information to recognize the context of the event. An RNN can read these sequences, but when a sequence is long, it forgets the earlier patterns of the sequence, and this difficulty is identified as the vanishing gradient. This problem can be resolved by using a specific kind of

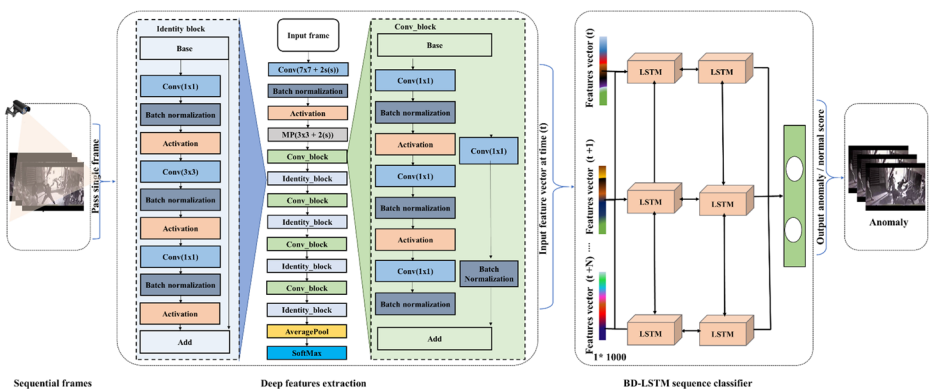
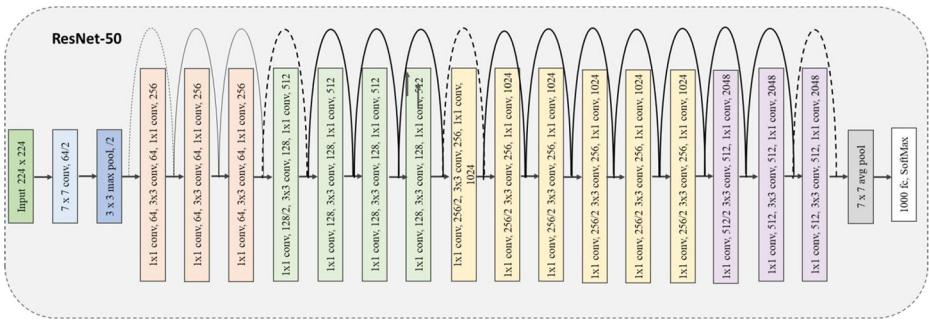


Fig. 1 The proposed anomaly detection framework for surveillance network



**Fig. 2** Architecture of ResNet-50 pre-trained on the ImageNet dataset. We extracted features from the last layer (fc\_1000) instead of using a SoftMax layer

RNN known as LSTM, due to its ability to keep the track of long sequences [3, 33]. Its chain-like building block with forget, input, and output gates, can regulate long-term sequence pattern recognition. The sigmoid function units are the main component of these gates, which acquires over the duration of training as it is open and closed. The dataflow and operations are processed by the LSTM unit from the input to the output gates. In Eqs. 1 to 7, “ $t$ ” is the input over time “ $t$ ”, the sigmoid function is represented by “ $\sigma$ ”, “ $\omega$ ” and “ $\beta$ ” are the weights and bias terms of the training stage, respectively [28]. “ $i_t$ ” “ $f_t$ ” and “ $O_t$ ” represent the three main gates of the LSTM at time “ $t$ ” (input, forget and output respectively). The input gate “ $i_t$ ” monitors when to record the current input “ $x_t$ ”, and the forget gate “ $f_t$ ” determines when to release the preceding memory cell “ $C_{t-1}$ ”. The output gate “ $O_t$ ” verifies the data shifted from current memory “ $C_t$ ” to the hidden state. In this state, the step is analysed using “ $\tanh$ ” activation and a memory cell “ $C_t$ ”. As anomaly detection does not require the transitional output of the LSTM, the final determination is made through the SoftMax classifier in the prediction state of the RNN.

$$i_t = \sigma(\omega_i[X_t^t + \zeta_{t-1}] + \beta_i)$$

$$f_t = \sigma(\omega_f[X_t^t + \zeta_{t-1}] + \beta_f)$$

$$\bar{O}_t = \sigma(\omega_o[X_t^t + \zeta_{t-1}] + \beta_o)$$

$$R = \tanh(\omega_r[X_t^t + \zeta_{t-1}] + \beta_r)$$

$$\hat{C}_t = \hat{C}_{t-1} \cdot f_t + R \cdot i_t$$

$$\zeta_t = \tanh(\hat{C}_t) \cdot \bar{O}_t$$

$$Prediction_{state} = SoftMax(V_{S_t})$$

The single LSTM cell cannot classify large amounts of training data such as the dynamic sequence patterns in videos. We therefore created a multi-layer-LSTM by assembling several LSTM cells to efficiently understand long-term sequence dependencies. The internal structure of an LSTM is presented in Fig. 3.

### 3.3 Multi-layer LSTM

The efficiency of a deep neural network depends upon the number of layers. A similar approach is pursued here for an RNN by stacking our network with two LSTMs, resulting in the learning of high-level sequence information [28]. In a typical LSTM for activation and processing, the data is transferred to a single layer prior to output however, in time sequence problems, we require to analyse information across multiple layers. Each and every layer in the LSTM is often a hierarchy that accepts the hidden state of the preceding layer as input via multiple LSTM layers, as shown in Fig. 4. The first layer of the LSTM, receives input sequential information as  $l$ , whereas the given input to the second layer is received from the preceding time step  $S_{t-1}$  and the output from the first layer is  $S_t$ . The LSTM cell calculation is similar to Eqs. 1 to 7 except that the information from each layer,  $i_t, f_t, O_t, C_t$  and  $S_t$ , is applied to every other layer. The process used to calculate the layer state is shown in Eq. 8.

$$S_t^l = tanh(\hat{C}_t^l) \cdot \bar{O}_t^l$$

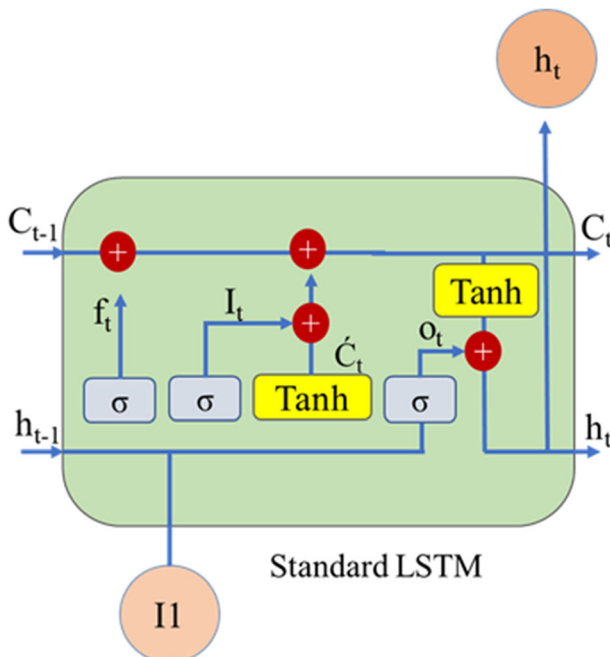


Fig. 3 Internal structure of a standard LSTM unit



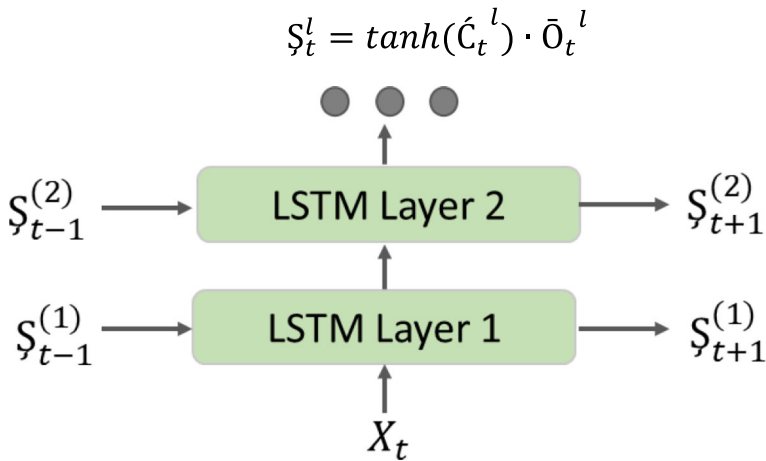


Fig. 4 Basic dataflow structure of a two-layer LSTM

### 3.4 Bi-directional LSTM

Unlike a simple LSTM, the output of the BD-LSTM depends not only on the previous frames but also on the upcoming frames in the sequence. The structure of the BD-RNN is relatively simple with two stacked RNNs, one in backwards direction and the other in a forward direction, while the hidden states of both RNNs are combined in the output. In this work, we used a multi-layer BD-LSTM in which each layer has two cells, one for the backwards pass and the other for the forward pass. The features

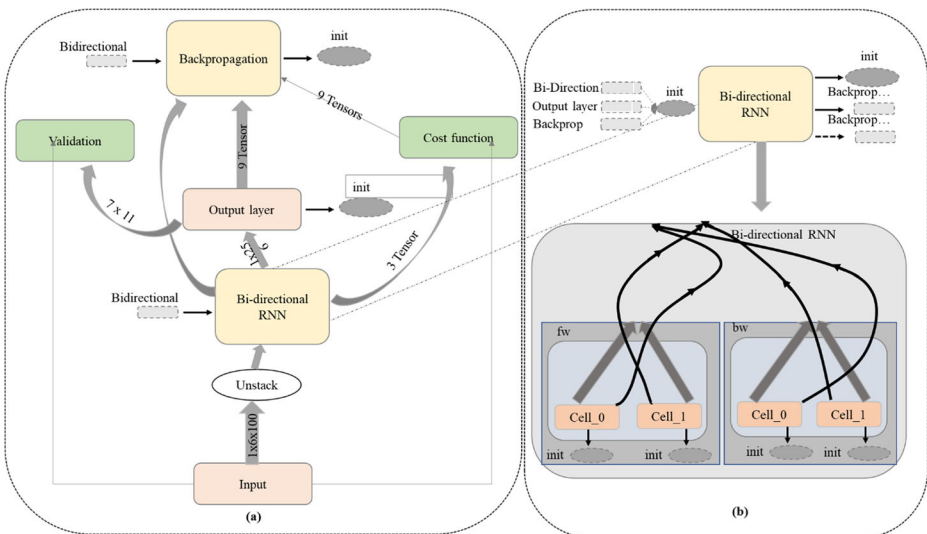


Fig. 5 The BD-LSTM network utilized in the proposed framework. **a** external structure and **b** internal structure of bidirectional the BD-LSTM. The top rightward indicates the flow of sequential data in the external structure of the BD-RNN cell

extracted by the ResNet-50 are used to decide either the normal or anomalous events that are fed to the multi-layer BD-LSTM in the form of chunks for an anomaly detection decision. The first chunk of the 1000 features from the initial frame of the video forms the input to the multi-layer BD-LSTM at time “ $t$ ”, while the next feature’s chunk at  $t + 1$  forms the second input to the multi-layer BD-LSTM, and so on. The overall structural design of the proposed multi-layer BD-LSTM is illustrated in Fig. 5. Figure 5a shows the training stage, in which the training data are passed to the model. The hidden state combines the forward and backward passes in the output layer, while backpropagation is used to adjust the bias and weights. A sample of 20% of the total data is used for validation purposes, while cross-entropy is employed for error rate assessment along with a learning rate as default of 0.001 using cost minimisation stochastic optimisation [22]. The interior structure of the BD-RNN is shown in Fig. 5b, where the forward and backward passes are represented as “fw” and “bw”, respectively. The method of computing the output allows the proposed model to accomplish state-of-the-art accuracy. The output of the frame is calculated based on the previous and upcoming frames, since each layer performs processing in both directions.

## 4 Experimental results and discussion

The proposed anomaly detection method is experimentally assessed utilizing two real-world large-scale anomaly video benchmark datasets including UCF-Crime [31] and UCFCrime2Local [17]. We used evaluation metrics that are typically applied in state-of-the-art schemes: the frame-based area under the curve (AUC) and the receiver operating characteristics (ROC) [18]. The proposed method was implemented in Python version 3.6 programming environment with a TensorFlow backend on a system with a GeForce-Titan-X graphics processing unit. The experimental results reveal the success of our proposed framework, as it detects anomalous events with greater precision than existing alternatives.

**Table 2** Statistical details of the UCF-Crime dataset

Type of anomaly	No of videos	Training set	Test set
Abuse	50	48	2
Arrest	50	45	5
Arson	50	41	9
Assault	50	47	3
Explosion	50	29	21
Fighting	50	45	5
Shooting	50	27	23
Shoplifting	50	29	21
Vandalism	50	45	5
Burglary	100	87	13
Stealing	100	95	5
Road accident	150	127	23
Robbery	150	145	5
Total	950	810	140

**Table 3** Statistical details of the UCFCrime2Local dataset

Attribute	Normal events	Anomalous events
Total number of videos	200	100
Training	141	69
Total length (minutes)	112.1	66.3
Min/max length (seconds)	4.6/ 135.9	6.8/59.8
Average length	39.8	33.6

#### 4.1 Anomaly detection datasets

Existing datasets for video anomaly detection has typically either limited classes of anomalies or contain a relatively small number of samples [18, 20, 26]. The UCF-Crime [31] and UCFCrime2Local datasets [17] are two largest and most diverse anomaly datasets. They contain a reasonable number of challenging surveillance real world anomaly videos and are suitable for analysing the efficiency and capabilities of our proposed method in terms of surveillance monitoring.

#### 4.2 UCF-crime dataset

The UCF-Crime dataset contains both normal and anomalous events videos, the latter of which contain 13 different kinds of anomalies, for example fighting, explosions, abuse and accidents, etc. The dataset comprises of 1900 surveillance videos with approximately equal number of normal and anomalous videos. The training portion of the dataset consisted of 800 normal and 810 anomalous samples, while the testing set included the remaining 150 normal and 140 anomalous videos. A summary of the UCF-Crime dataset is given in Table 2.

#### 4.3 UCFCrime2Local dataset

The UCFCrime2Local dataset consists of six human-based classes of anomalies: burglary, arrest, robbery, assault, stealing and vandalism. This dataset contains a total of 300 videos with 100 anomalous and 200 normal videos. The dataset was divided into training and test sets with the training set containing 210 and the test set have 90 videos. We tested our proposed framework using the annotation test video provided in [31] in which a weakly supervised

**Table 4** Comparative analysis based on accuracy for CNN features integrated with a variety of sequential models

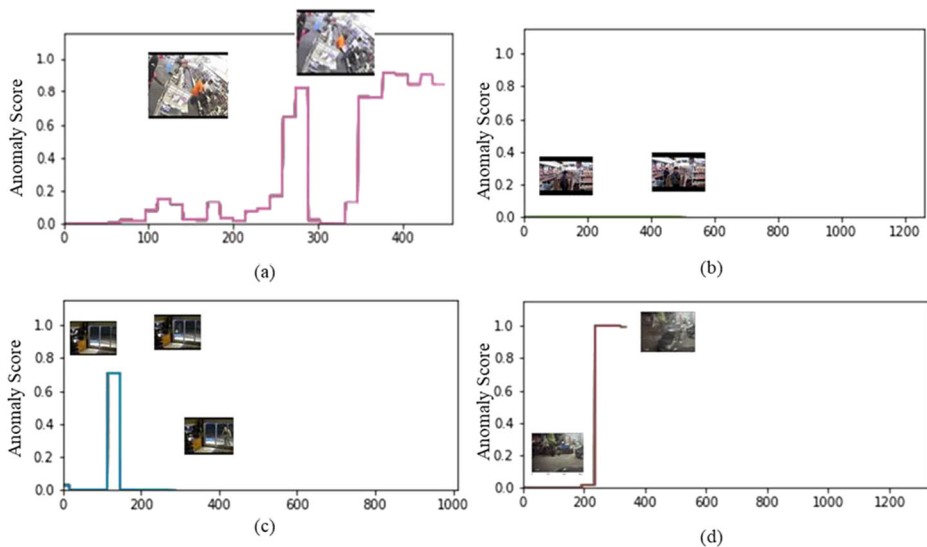
Pre-trained Feature Extraction Model	Sequence Learning Model	Accuracy (%)	
		UCF-Crime	UCFCrime2Local
VGG-19 [30]	RNN	78	77.5
	LSTM	80	86.2
	Multi-layer BD-LSTM	82	87.5
Inception V3 [32]	RNN	71	70
	LSTM	79	77
ResNet-50 [12]	Multi-layer BD-LSTM	80	88
	RNN	78	82
	LSTM	84	88
	Multi-layer BD-LSTM	85.53	89.05

**Table 5** Comparative analysis of the proposed method with state-of-the-art techniques based on AUC

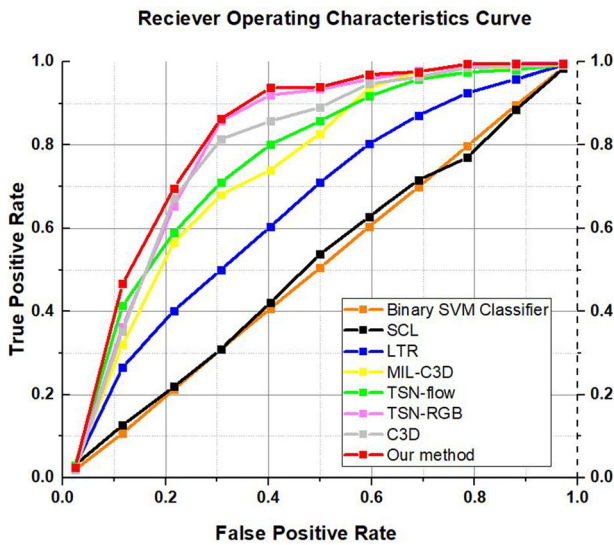
Methods	AUC (%)	
	UCF-Crime dataset [31]	UCFCrime2Local dataset [17]
Binary SVM classifier [31]	50.0	–
LTR [11]	50.6	–
SCL [20]	65.51	–
MIL-C3D. without constraints [31]	74.44	–
MIL-C3D. with constraints [31]	75.41	–
C3D [40]	81.08	–
TSN-RGB [40]	82.12	–
TSN-optical flow [40]	78.08	–
Video segment [17]	–	56.12
Spatiotemporal [10]	63	68
Oracle tube [17]	–	74.73
Weakly supervised [17]	–	80.96
<b>Proposed method</b>	<b>85.53</b>	<b>89.05</b>

The bold text demonstrates the best performance achieved by the proposed method on both UCF-Crime and UCFCrime2Local datasets

method was used to train the paradigm. The statistical details of UCFCrime2Local and a comparative analysis of the various CNN features are presented in Tables 3 and 4. We performed experiments on the features of various CNNs by integrating our approach with different sequential models. The VGG-19 with multi-layer BD-LSTM achieved 82% accuracy on UCF-Crime and 87.5% on UCFCrime2Local, while the inception V3 with multi-layer BD-LSTM reached 80% accuracy on UCF-Crime and 88% on UCFCrime2Local. The ResNet-50 with multi-layer BD-LSTM achieved greater success, with an accuracy of 85.53% on UCF-Crime and 89.05% on UCFCrime2Local.



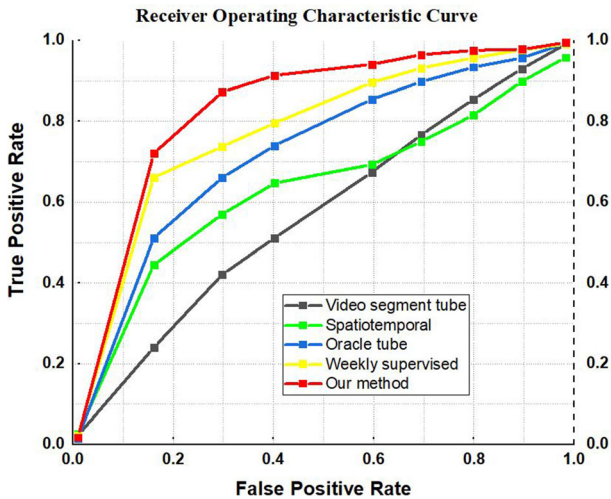
**Fig. 6** Qualitative results from our framework on sample testing videos. **a** anomalous shoplifting activity; **b** normal activities; **c** vandalism; and **d** a road accident



**Fig. 7** Comparison of the proposed method with existing techniques using the UCF-Crime dataset. ROC results for binary SVM classifier [31], SCL [11], LTR [20], MIL-C3D [31], TSN-flow [40], TSN-RGB [40], C3D [40], and our framework are represented by orange, black, blue, yellow, green, pink, grey, and red colours, respectively

**4.4 Results**

In this section, our proposed framework is experimentally assessed utilizing various datasets, and the results are shown in Table 5. We compare our framework with other existing techniques [10, 17, 31, 40, 41] for anomaly detection. The evaluation metrics used to check our framework are the ROC and AUC. Some visual results from our framework for anomalous frame detection are shown in Fig. 6.



**Fig. 8** Comparison of the proposed method with existing techniques for the UCFCrime2Local dataset. ROC comparison of video segment tube [17], spatiotemporal method [10], Oracle Tube [17], a weakly supervised method [17], and our framework are represented by black, green, blue, yellow, and red colours, respectively

**Table 6** Comparison of the proposed method with state-of-the-art in terms of the parameters, model size, and time complexity

Model	No. of parameters (million)	No. of weighted layers/blocks	Model size (MBs)	Time complexity/ Per sequence
C3D [31]	–	11	313	–
VGG-19+ multi-layer BD-LSTM [30]	143	19	605.5	0.22
Inception V3+ multi-layer BD-LSTM [32]	23	42	148.5	–
ResNet-50 + multi-layer BD-LSTM [12]	25	50	143	0.20

#### 4.5 Comparison with the state-of-the-art techniques

We compared our anomaly detection method with other existing alternatives for the UCF-Crime and UCFCrime2Local datasets. We used only the test set from the UCFCrime2Local dataset to establish the AUC and ROC to enable a comparison with the results in [10, 17]. In [10], the authors used weakly supervised spatiotemporal features with a MIL scheme for anomaly detection, while the authors of [17] used various approaches including video segment, the Oracle Tube, and a weakly supervised spatiotemporal tube for anomaly locality in videos. The video segment approach is not fully supervised, and the Oracle Tube is based on an annotated video supervised approach. The experimental results presented in Figs. 7 and 8 using ROC demonstrate that our framework accomplishes better performance than the approaches in [10, 11, 17, 20, 31, 40]. The AUC scores are compared in Table 5, and it can be noticed that the our framework achieved the highest AUC of 85.53%, an increase of 3.41% compared to the method of Zhong et al. [40], which had the next greatest AUC of 82.12%. Some other followed researches in [11, 20, 31] achieved AUC scores of 75.41%, 50.6% and 65.51%, respectively, proving that our framework gives good performance for both datasets. The proposed deep model processes a sequence of 15 frames within 0.20 s. We compared our proposed model with the famous C3D [31] model, which utilised a massive number of parameters and 3D filters, meaning that the size of the trained model was large compared to the proposed model. We also performed experiments with the VGG-19 and inception V3 models. The size of the VGG-19 model was too big, and its real-time performance was lower than the ResNet-50 with LSTM. Similarly, the inception V3 model used fewer parameters, after integration with the multi-layer BD-LSTM, the size of the model was larger than the ResNet-50 with LSTM. After an assessment of the other existing alternatives, we found that our model processed a single second video in half a second, with no sign of delay, proving that

**Table 7** Comparison of the proposed framework with recent state-of-the-art techniques based on false alarm rate

Methods	UCF-Crime dataset [31]	UCFCrime2Local dataset [17]
MIL-C3D with constraints [31]	1.9	–
Hasan et al. [11]	27.2	–
Lu et al. [20]	3.1	–
C3D [40]	2.8	–
TSN-RGB [40]	0.1	–
TSN-optical flow [40]	1.1	–
Proposed method	0.44	0.71

our framework is suitable for implementation in real-time scenarios. Table 6 shows the time complexity and model size comparison of the proposed framework with other deep learning models. Furthermore, the frequency of normal events is greater than anomalous events in daily life surveillance, thereby a robust and effective anomaly detection systems with minimal false alarm rates is demand of the time. Therefore, we compared the proposed framework with recent existing approaches as given in Table 7 where the statistics indicate lowest false alarm rates of the proposed anomaly detection framework.

## 5 Conclusion

In this study, we presented an efficient framework for real-world anomaly detection in surveillance environments with state-of-the-art accuracy on existing anomaly detection datasets. The generic pipeline of our framework extracted deep CNN features from sequential frames, followed by a new multi-layer BD-LSTM for normal and anomaly class detection. The use of deep features combined with the multi-layer BD-LSTM gives a high-level of adaptability in terms of training and validation data and is applicable to real-world surveillance networks. The proposed framework is demonstrated to have higher accuracy in comparison with anomaly detection methods in the recent literature. The experimental results indicate an increase of 3.41% for the UCF-Crime dataset and 8.09% for the UCFCrime2Local dataset. Currently, the accuracy of our framework is inadequate for low disparity and needs further improvements, particularly since the UCF-Crime dataset involves very challenging categories. In future, we have intention to investigate motion features, strong visual features, and two stream CNN networks for anomaly detection to overcome the challenge of lower variation, which is not efficiently detected by our current framework.

**Acknowledgements** “This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (2019-0-00136, Development of AI-Convergence Technologies for Smart City Industry Productivity Innovation).”

## References

1. Adam A, Rivlin E, Shimshoni I, Reinitz D (2008) Robust real-time unusual event detection using multiple fixed-location monitors. *IEEE Trans Pattern Anal Mach Intell* 30:555–560
2. Al Ridhawi I, Otoum S, Aloqaily M, Jararweh Y, Baker T (2020) Providing secure and reliable communication for next generation networks in smart cities. *Sustain Cities Soc* 56:102080
3. Al-Smadi M, Qawasmeh O, Al-Ayyoub M, Jararweh Y, Gupta B (2018) Deep recurrent neural network vs. support vector machine for aspect-based sentiment analysis of Arabic hotels’ reviews. *J Comput Sci* 27: 386–393
4. Alsmirat MA, Obaidat I, Jararweh Y, Al-Saleh M (2017) A security framework for cloud-based video surveillance system. *Multimed Tools Appl* 76:22787–22802
5. Benezeth Y, Jodoin P-M, Saligrama V, Rosenberger C (2009) Abnormal events detection based on spatio-temporal co-occurrences. In: 2009 IEEE conference on computer vision and pattern recognition, pp 2458–2465
6. Cheng K-W, Chen Y-T, Fang W-H (2015) Gaussian process regression-based video anomaly detection and localization with hierarchical feature representation. *IEEE Trans Image Process* 24:5288–5301
7. Cheng K-W, Chen Y-T, Fang W-H (2016) An efficient subsequence search for video anomaly detection and localization. *Multimed Tools Appl* 75:15101–15122

8. Chong YS, Tay YH (2017) Abnormal event detection in videos using spatiotemporal autoencoder. In: International symposium on neural networks, pp 189–196
9. Cong Y, Yuan J, Liu J (2011) Sparse reconstruction cost for abnormal event detection. In: CVPR 2011, pp 3449–3456
10. Gianchandani U, Tirupattur P, Shah M Weakly-supervised spatiotemporal anomaly detection
11. Hasan M, Choi J, Neumann J, Roy-Chowdhury AK, Davis LS (2016) Learning temporal regularity in video sequences. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 733–742
12. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778
13. He C, Shao J, Sun J (2018) An anomaly-introduced learning method for abnormal event detection. *Multimed Tools Appl* 77:29573–29588
14. Huo J, Gao Y, Yang W, Yin H (2012) Abnormal event detection via multi-instance dictionary learning. In: International conference on intelligent data engineering and automated learning, pp 76–83
15. Kim J, Grauman K (2009) Observe locally, infer globally: a space-time MRF for detecting abnormal activities with incremental updates. In: 2009 IEEE conference on computer vision and pattern recognition, pp 2921–2928
16. Kratz L, Nishino K (2009) Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models. In: 2009 IEEE conference on computer vision and pattern recognition, pp 1446–1453
17. Landi F, Snoek CG, Cucchiara R (2019) Anomaly locality in video surveillance. arXiv preprint arXiv:1901.10364
18. Li W, Mahadevan V, Vasconcelos N (2013) Anomaly detection and localization in crowded scenes. *IEEE Trans Pattern Anal Mach Intell* 36:18–32
19. Liu W, Luo W, Lian D, Gao S (2018) Future frame prediction for anomaly detection—a new baseline. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 6536–6545
20. Lu C, Shi J, Jia J (2013) Abnormal event detection at 150 fps in matlab. In: Proceedings of the IEEE international conference on computer vision, pp 2720–2727
21. Luo W, Liu W, Gao S (2017) A revisit of sparse coding based anomaly detection in stacked rnn framework. In: Proceedings of the IEEE international conference on computer vision, pp 341–349
22. Luo W, Liu W, Gao S (2017) Remembering history with convolutional lstm for anomaly detection. In: 2017 IEEE International Conference on Multimedia and Expo (ICME), pp 439–444
23. Mahadevan V, Li W, Bhalodia V, Vasconcelos N (2010) Anomaly detection in crowded scenes. In: 2010 IEEE computer society conference on computer vision and pattern recognition, pp 1975–1981
24. Mehran R, Oyama A, Shah M (2009) Abnormal crowd behavior detection using social force model. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp 935–942
25. Muhammad K, Hussain T, Tanveer M, Sannino G, de Albuquerque VHC (May 2020) Cost-effective video summarization using deep CNN with hierarchical weighted fusion for IoT surveillance networks. *IEEE Internet Things J* 7:4455–4463
26. Rabiee H, Haddadnia J, Mousavi H, Kalantarzadeh M, Nabi M, Murino V (2016) Novel dataset for fine-grained abnormal behavior understanding in crowd. In: 2016 13th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pp 95–101
27. Sabokrou M, Fayyaz M, Fathy M, Klette R (2017) Deep-cascade: cascading 3d deep neural networks for fast anomaly detection and localization in crowded scenes. *IEEE Trans Image Process* 26:1992–2004
28. Sak H, Senior A, Beaufays F (2014) Long short-term memory recurrent neural network architectures for large scale acoustic modeling. In: Fifteenth annual conference of the international speech communication association
29. Shao J, Loy C-C, Kang K, Wang X (2016) Slicing convolutional neural network for crowd video understanding. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 5620–5628
30. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556
31. Sultani W, Chen C, Shah M (2018) Real-world anomaly detection in surveillance videos. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 6479–6488
32. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z (2016) Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2818–2826
33. Ullah A, Ahmad J, Muhammad K, Sajjad M, Baik SW (2017) Action recognition in video sequences using deep bi-directional LSTM with CNN features. *IEEE Access* 6:1155–1166
34. Wang T, Qiao M, Zhu A, Niu Y, Li C, Snoussi H (2018) Abnormal event detection via covariance matrix for optical flow based feature. *Multimed Tools Appl* 77:17375–17395



35. Xu D, Yan Y, Ricci E, Sebe N (2017) Detecting anomalous events in videos by learning deep representations of appearance and motion. *Comput Vis Image Underst* 156:117–127
36. Yu Y, Zhao T, Wang M, Wang K, He L (2020) Uni-OPU: an FPGA-based uniform accelerator for convolutional and transposed convolutional networks. In: *IEEE transactions on very large scale integration (VLSI) systems*
37. Zhang T, Jia W, Yang B, Yang J, He X, Zheng Z (2017) MoWLD: a robust motion image descriptor for violence detection. *Multimed Tools Appl* 76:1419–1438
38. Zhang J, Kalantidis Y, Rohrbach M, Paluri M, Elgammal A, Elhoseiny M (2019) Large-scale visual relationship understanding. In: *Proceedings of the AAAI conference on artificial intelligence*, pp 9185–9194
39. Zhao B, Fei-Fei L, Xing EP (2011) Online detection of unusual events in videos via dynamic sparse coding. In: *CVPR 2011*, pp 3313–3320
40. Zhong J-X, Li N, Kong W, Liu S, Li TH, Li G (2019) Graph convolutional label noise cleaner: train a plug-and-play action classifier for anomaly detection. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp 1237–1246
41. Zhu Y, Newsam S (2019) Motion-aware feature for improved video anomaly detection. *arXiv preprint arXiv:1907.10211*

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