Spatiotemporal Anomaly Detection Using Deep Learning for Real-Time Video Surveillance

Rashmika Nawaratne , Damminda Alahakoon, *Member, IEEE*,
Daswin De Silva , *Member, IEEE*, and Xinghuo Yu , *Fellow, IEEE*

Abstract—**Rapid developments in urbanization and autonomous industrial environments have augmented and expedited the need for intelligent real-time video surveillance. Recent developments in artificial intelligence for anomaly detection in video surveillance only address some of the challenges, largely overlooking the evolving nature of anomalous behaviors over time. Tightly coupled dependence on a known normality training dataset and sparse evaluation based on reconstruction error are further limitations. In this article, we propose the incremental spatiotemporal learner (ISTL) to address challenges and limitations of anomaly detection and localization for real-time video surveillance. ISTL is an unsupervised deep-learning approach that utilizes active learning with fuzzy aggregation, to continuously update and distinguish between new anomalies and normality that evolve over time. ISTL is demonstrated and evaluated on accuracy, robustness, computational overhead as well as contextual indicators, using three benchmark datasets. Results of these experiments validate our contribution and confirm its suitability for real-time video surveillance.**

Index Terms—**Active learning, anomaly detection, anomaly localization, deep learning, real-time video surveillance, spatiotemporal analysis, unsupervised learning.**

I. INTRODUCTION

VIDEO surveillance is a predominant consideration in the development, operation, and sustainability of modern industrial and urban environments. It contributes toward efficiency, safety, security, and optimality of the locality, infrastructure, individuals, operations, and activities [1]. Industrial environments are transitioning toward autonomous machinery, cyber-physical systems, and energy-efficient layouts. Urban environments are becoming densely populated, with high usage of multilevel buildings, increased vehicular, pedestrian, and crowd movements. This vertical and horizontal expansion of asset and area utilization in both industrial and urban environments has eventuated an exponential increase in the deployment of closed-circuit television camera systems [1]. However, it is unrealistic and infeasible for human observers to monitor and analyze every video stream with high precision. Artificial intelligence (AI) techniques for autonomous video surveillance reported in current literature can be categorized into video summarization [2]–[4], object detection and re-identification [5]–[7], activity/behavior detection [8]–[10], and anomaly detection [11], [12].

Anomaly detection is a constitutive task in autonomous video surveillance as it contributes to the success of the other categories noted above. It is also a complex task as the anomalies to be detected are not known prior, imposing difficulties even for a human observer. A general definition of anomaly detection is the identification of behaviors that do not conform to expected and accepted behavior (i.e., normal behavior) [13]. In the context of autonomous video surveillance, anomaly detection is impacted by three primary challenges. First, the computational complexity and cost of video data processing due to spatial and temporal dimensional structure combined with nonlocal temporal variations across video frames [11], [14], [15]. As an example, anomalous objects such as vehicles/bicycles in a pedestrian walk must be identified using spatial processing, whereas anomalous behavior such as jaywalking must be determined using temporal variations across video frames. Second, the anomaly itself is ill-defined, the boundary between normal behavior and anomalies is often imprecise, and anomalies are highly contextual [13]. For example, industrial machinery operating at low power can be either normal or anomalous depending on the operational circumstances. Third, what is considered as normal behavior evolves over time making current knowledge incomplete and/or obsolete [15], [16]. For instance, when offenders become aware of detected anomalies, they can maliciously adapt behaviors so that subsequent anomalies are difficult to detect.

Existing literature attempts to address the first and second challenges (i.e., computational complexity and identifying contextual anomalies). The third challenge, the evolving nature of normal behavior over time, remains unaddressed, and this makes current knowledge of normality incomplete. In this article, we propose the incremental spatiotemporal learner (ISTL) to address the aforementioned challenges and limitations. ISTL is a new anomaly detection approach for real-time video surveillance that actively learns spatiotemporal patterns of normal

behavior as it evolves over time. ISTL is inspired by continuous learning process in human cognition and the paradigm of active learning. Inspired by the human brain, ISTL begins by developing a basic understanding from immediately available information to distinguish between normal (safe) and anomalous (unsafe) behaviors and continuously refines this understanding as the surroundings change and new information becomes available [17]. Active learning is primarily used for refinement and validation in ISTL, where a human observer contributes to the learning process for improved learning outcomes across iterations. The paradigm of active learning has been widely used in industrial image and video analysis applications such as character reading, facial recognition, autonomous vehicles, and e-commerce [18], [19].

The research contributions of this research article are as follows.

1) A deep learning model that learns spatiotemporal patterns of normal behavior for online anomaly detection and localization from a surveillance video stream.

2) The fuzzy aggregation of active learning outcomes into the continuous learning process for dynamic adaption to evolving behaviors of unknown/new normalities in the surveillance video stream.

3) Evaluation using two thresholds, anomaly threshold and temporal threshold, based on the context of the video surveillance feed, instead of sparse evaluation based solely on reconstruction error. Key features of ISTL are demonstrated and validated using three benchmark video surveillance datasets: University of California San Diego (UCSD) Pedestrian datasets [20] (Ped 1 and Ped 2) and The Chinese University of Hong Kong (CUHK) Avenue dataset [21].

The rest of this article is organized as follows. Section II reports related work. Section III presents the proposed ISTL approach, with descriptions of each phase and corresponding outcomes. Section IV presents evaluation of the ISTL approach for accuracy, robustness, low computational overhead, and contextual indicators, across a range of values/scenarios for both anomaly and temporal thresholds. Section V concludes this article.

II. RELATED WORK

Techniques and approaches for intelligent video surveillance in current literature broadly range across two areas of research, hand-crafted video features [8], [22], [23], and learned-representations based on deep learning architectures [11], [12], [24], [25]. In techniques that utilize hand-crafted features, trajectories and spatiotemporal changes are extracted as input/output features for computational and AI modeling. For instance, Xie and Guan [26] proposed a motion instability based anomaly detection framework that discriminates anomalous behavior based on the direction randomness and motion intensity, whereas Wu et al. [27] proposed an approach in which objects are classified as anomalous based on how they follow the learned normal trajectory. These trajectory-based methods define normal behavior based on previously observed motion patterns. However, such trajectory-based methods fail to detect

anomalous behavior based on the appearance of entities in the surveillance video stream and computationally expensive for crowded scenes. State-of-the-art handcrafted feature extraction methods describe video events ranging from pixel-level to three-dimensional (3-D) cuboid. For instance, Zhao et al. [28] utilize histogram of gradient and histogram of optical flow along spatial and temporal dimensions to encode an event and learn the normality upon dynamic sparse coding, whereas Zaharescu and Wildes [29] models the normal behavior based on distributions of spatiotemporal oriented energy. These handcrafted feature-based techniques can accurately model both spatial and temporal dynamics, however, they require prior knowledge for the design of effective features, and are time consuming to extract, thereby impractical to use in real-time anomaly detection.

With the advancements of deep learning, convolutional neural networks (CNN), autoencoders and recurrent neural networks (RNN) have been utilized for video anomaly detection [24]. Xu et al. [12] proposed Appearance and Motion DeepNet (AMDN) that utilizes an autoencoder to automatically learn feature representation from the surveillance video, use a double fusion framework and support vector machine models to predict the irregularity of an event. The AMDN model results in the state-of-the-art accuracy, however, its processing time is in the order of 10 000 ms, which makes it impractical to use in online anomaly detection. Hasan et al. [25] approached the problem of anomaly detection by learning a generative model for regular motion patterns. The approach achieved positive results using a ten-layered fully convolutional feed-forward autoencoder to reconstruct input video, then detect anomalies based on its reconstruction cost analysis. Luo et al. [30] attempted to detect anomalies by leveraging a CNN for appearance encoding and a convolutional long short-term memory (ConvLSTM) for remembering history of the motion information. Recently, Vu et al. [31] proposed an anomaly detection approach using a deep generative network in which normality is modeled by an unsupervised probabilistic framework. With these advancements, it is evident that learned-representations based on deep learning architectures have the ability to distinguish anomalies from normal behaviors by processing high-dimensional surveillance video streams. However, existing deep learning approaches for anomaly detection are highly dependent on a known normality dataset for training and constrained by sparse evaluation based only on reconstruction error, without consideration for surveillance context.

In summary, current literature is mostly limited to addressing the computational complexity of processing high-dimensional video data and identifying contextual anomalies from surveillance video streams. To the best of our knowledge, the challenge of evolving nature of normal behavior over time remains unaddressed which makes current knowledge of normality incomplete.

III. ISTL APPROACH

A high-level overview of ISTL is illustrated in Fig. 1. First, the live video surveillance feed is presented as input to spatiotemporal model training of normal behavior (from time t_0 to t_u). Second, the trained model is utilized for anomaly detection
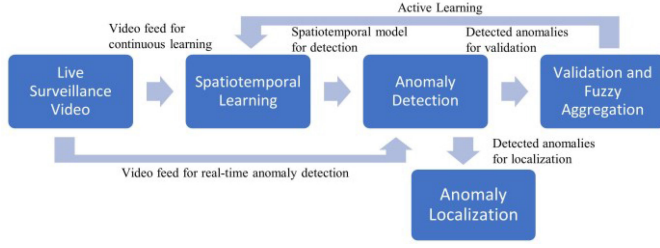
Fig. 1. Overview of the proposed ISTL approach.

and localization within the time interval t_u to t_v. Third, the detected anomalies are validated by a human observer and the validation input is used to construct updated normal behavior using fuzzy aggregation. This updated normal behavior is fed back into the ISTL learning model for continuous learning.

This overview is expanded into a functional view and illustrated in Fig. 2.

The computational formulation of anomaly detection in video surveillance is presented as follows. The training video stream (X_train) composed of a sequence of frames of height h and width w, $X_\text{train} \subset R$, that only contains video frames exhibiting normal behavior in a given camera view. R indicates all the video frames of the camera view in real world. In the testing phase, a video stream (X_test) is employed, where $X_\text{test} \subset R$ contains video frames of both normal and anomalous behavior. The goal is to learn a representation (Ω) of normal behavior from X_train which is subsequently validated with X_test to distinguish anomalies. In contrast to previous work [25], [32]–[34], which requires a complete training dataset of normal behavior, the ISTL approach will actively update previously learned knowledge (Ω) based on 1) spatiotemporal information from continuously received video streams and 2) active human observer feedback on detected anomalies.

The three phases of ISTL, spatiotemporal learning, anomaly detection and localization, active learning with fuzzy aggregation, are explicated in following subsections.

A. Spatiotemporal Learning

Spatiotemporal representation of normal behavior is learned from X_train as expected and acceptable behavior for the video surveillance application. The ISTL model is composed of a spatiotemporal autoencoder to learn the appearance and motion representation from video inputs. The autoencoder is an unsupervised learning algorithm that employs backpropagation to set the target values to be equal to the inputs by minimizing the reconstruction error [35]. In the proposed architecture, the spatiotemporal autoencoder consists of a series of CNN layers to learn the spatial representation and a series of ConvLSTM layers to learn the temporal representation. The input data layer and feature transformation layers of the autoencoder are described in the following sections.

1) Input Data Layer: The raw video data are preprocessed to enhance the learning capacity of the spatiotemporal autoencoder model. At first, the video data are extracted as consecutive frames, convert into grayscale to reduce the dimensions, resize

to 224×224 pixels and normalize pixel values by scaling between 0 and 1. The input to the spatiotemporal autoencoder model is a temporal cuboid of video frames, which will be extracted using a sliding window of length T without any feature transformation. The consecutive frames of length T are stacked together to construct the input temporal cuboid. Increased length of this temporal window (T) will enable to incorporate motion of longer length, however, the larger the T, the model convergence will take exponential time [25].

2) Convolution Layers (CNN): CNNs have been inspired by biological processes resembling the organization of the animal visual cortex [36]. The connectivity of the neurons in the convolution layers is designed in a manner similar to animal vision system such that an individual cortical neuron responds to stimuli only in a confined region of the input frame, i.e., the receptive field. In video analysis, the convolution layers can preserve the spatial relationship within the input frames by learning feature representations using filters, whose values are learned during the training process. The ISTL model consists of two convolution layers and two de-convolution layers, whose filters and kernel sizes are specified in Table I.

3) ConvLSTM Layers: RNN captures the dynamic temporal behavior of a time-sequence input data by employing an internal memory to process the input sequences. LSTM units are advancement of generic building blocks of the RNN. The LSTM unit is composed of an input gate, an output gate, a forget gate, and a cell. The input gate defines the extent the input value moves into the unit. The forget gate controls the extent the values from the previous time steps remain in the unit and the output gate controls to which extent the current input value is used for the computation for the activation of the unit. The cell remembers values over arbitrary time intervals.

As LSTM is primarily developed and utilized for modeling long-range temporal correlations, it has a drawback in handling spatial data as spatial information is not encoded in its state transition. However, it is essential to learn the temporal regularity from the surveillance video stream while preserving the spatial structure, particularly for anomaly detection. Therefore, we utilize an extension to LSTM, ConvLSTM [37], in which both the input-to-state and state-to-state transitions have convolution structures. The ConvLSTM overcomes this drawback by designing its inputs, hidden states, gates, and cell outputs as 3-D tensors, whose last dimension is the spatial dimension. Furthermore, the matrix operations in its inputs and gates are replaced with convolution operator. With these modifications, the ConvLSTM is able to capture the spatiotemporal features from the input frame sequences. The ConvLSTM model is represented as following:

$$i_t = \sigma\left(W_{xi} * X_t + W_{hi} * H_{t-1} + W_{ci} \circ C_{t-1} + b_i\right) \quad (1)$$

$$f_t = \sigma\left(W_{xf} * X_t + W_{hf} * H_{t-1} + W_{cf} \circ C_{t-1} + b_f\right) \quad (2)$$

$$C_t = f_t \circ C_{t-1} + i_t \circ \tanh\left(W_{xc} * X_t + W_{hc} * H_{t-1} + b_c\right) \quad (3)$$

$$o_t = \sigma\left(W_{xo} * X_t + W_{ho} * H_{t-1} + W_{co} \circ C_{t-1} + b_o\right) \quad (4)$$

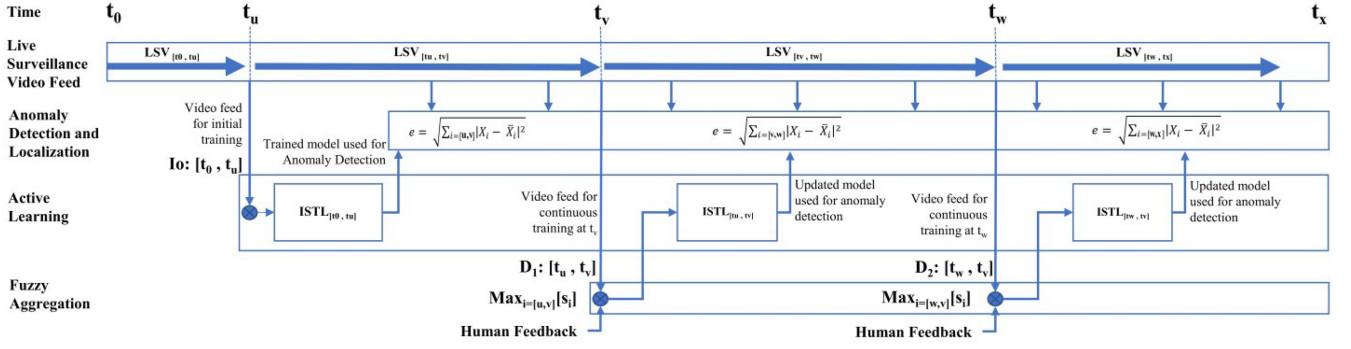$$H_t = o_t \circ \tanh\left(C_t\right). \quad (5)$$

Fig. 2. Functional view of the ISTL approach. *I*: input frames, *D*: detected frames, t_u: time step at initial training, t_v and t_w: time steps at second and third training iterations, *LSV*: Live surveillance video.

TABLE I
SPATIOTEMPORAL AUTOENCODER ARCHITECTURE

ID	Input Tensor	Operation	Output Tensor
C1	T*224*224*1	CV; F: 128; K: 27*27; S: 4	T*56*56*128
C2	T*56*56*128	CV; F: 64; K: 13*13; S: 2	T*28*28*64
CL1	T*28*28*64	CL; F: 64; K: 3*3	T*28*28*64
CL2	T*28*28*64	CL; F: 32; K: 3*3	T*28*28*32
DCL1	T*28*28*32	CL; F: 64; K: 3*3	T*28*28*64
DC1	T*28*28*64	DCV; F: 64; K: 13*13; S: 2	T*56*56*128
DC2	T*56*56*128	DCV; F: 128; K: 27*27; S: 4	T*224*224*1

[CV] Convolution, [CL] Convolution LSTM, [DCV] Deconvolution, [T] Depth of temporal cuboid, [F] Number of filters, [K] Kernel size, [S] strides, [*] Multiplication.
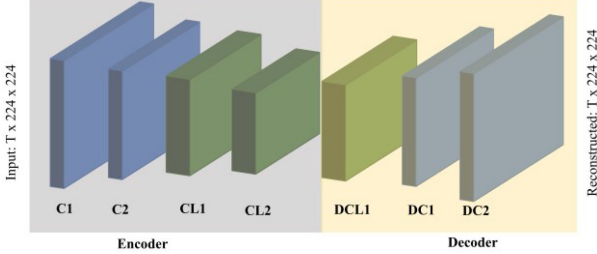

Fig. 4. Anomaly detection and localization.

Fig. 3. Spatiotemporal autoencoder architecture. Layer IDs are referred from Table I. Best viewed in color.

In the equations, "∗" and "∘" represents convolution operation and Hadamard product, respectively. Inputs are represented by $X_i, \ldots, X_t$, the cell states are represented by $C_i, \ldots, C_t$, the hidden states are represented by $H_i, \ldots, H_t$, and the gates i_t, f_t, and o_t are all 3-D tensors. "σ" is the sigmoid function and, $W_{x\sim}$ and $W_{h\sim}$ are two-dimensional (2-D) convolution kernels in the ConvLSTM. The ISTL model consists of three ConvLSTM layers. The spatiotemporal autoencoder architecture is illustrated in Fig. 3 and its composition further elaborated in Table I.

B. Anomaly Detection and Localization

The ISTL model can be used to obtain a reconstruction of the normality of the input video at pixel-level precision. However, the trained autoencoder does not have the ability to accurately reconstruct the anomalous or unseen scenes, due to the fact that such scenes have not been p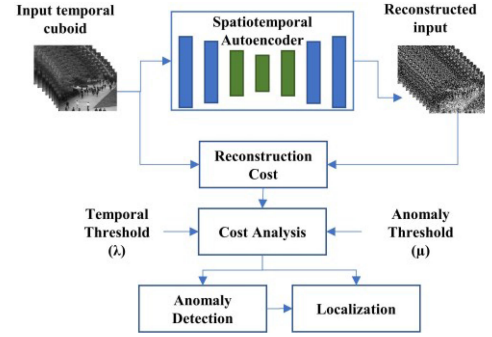resented in the training phase. This phenomenon is used to evaluate and detect anomalies from the input video. We obtain the reconstruction error (*E*) as the square root of the sum of the squared vector values, as represented in (6) and (7), where *X* is the input temporal cuboid, $\bar{X}$ is the reconstructed temporal cuboid, *T* is the time window, *w* is the width, and *h* is the height of the video frame

$$\varphi(i, j, k) = \left| X_{(i,j,k)} - \bar{X}_{(i,j,k)} \right|^2 \qquad (6)$$

$$E = \left(\sum_{i=0}^{T} \sum_{j=0}^{w} \sum_{k=0}^{h} \varphi(i, j, k) \right)^{\frac{1}{2}}. \qquad (7)$$

The reconstruction error represents the score for each temporal cuboid defining the anomaly. We define a reconstruction error threshold to distinguish between normal behavior and anomalies, named anomaly threshold (μ). In practical video surveillance applications, the human observer can select a value for μ based on the sensitivity required for the surveillance application. A low μ would result in higher sensitivity to the surveillance arena, resulting in a higher number of alerts. A high μ would result in lesser sensitivity that could lead to miss sensitive anomalies in the surveillance arena.

Additionally, we introduce the temporal threshold (λ), which we define as the number of video frames that should be higher than the μ to recognize an event as an anomaly. λ is employed to reduce the false-positive anomaly alerts due to sudden variations of the surveillance video stream due to occlusion, motion blur,
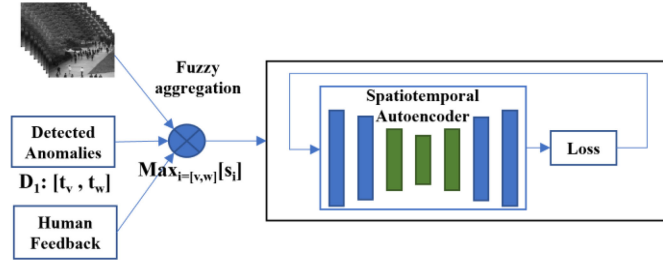
Fig. 5. Active learning of spatiotemporal autoencoder model. *D*: detected input frames, *v, w*: time epoch of previous anomaly detection.

and high-intense lighting conditions. Fig. 4 illustrates anomaly detection approach based on the reconstruction error.

Anomaly localization locates the specific area of the video frame, where an anomaly has occurred. Subsequent to detecting a segment of the video as anomalous, we localize the anomalies by calculating reconstruction error (E_c) over nonoverlapping spatiotemporal local cuboid windows, where m and n are the width and height, respectively, and T is the depth (i.e., number of frames in the cuboid). Equation (7) is used to calculate the E_c for local cuboids.

C. Active Learning With Fuzzy Aggregation

The purpose of the active learning in practical video surveillance context is to enable anomaly detection of dynamically evolving environments. By automating the anomaly detection using the deep learning model discussed in Section III-A and -B, we train the learning model to identify accepted normal behavior provided at the beginning. However, in dynamic environments comprising of new normal behavior that have not anticipated and/or existing behavior that considered abnormal reformed to normal, it is important that the detection system evolves with capabilities for detecting such new scenarios. ISTL addresses this challenge by adopting an active learning approach using fuzzy aggregation to continuously train the learning model with unknown/new normal behavior specific to the corresponding surveillance context. This approach is inspired by the human brain's ability to develop a basic understanding which is continuously refined as new information becomes available [17].

ISTL is initially trained with a preidentified normal behavior in the surveillance context and used for anomaly detection. If a video frame is detected as an anomaly, i.e., the reconstruction error of the input cuboid is above the anomaly threshold, the input cuboid is classified as an anomaly. The classified frames are then sent to a human observer for verification. The objective of human observer feedback is to actively feed the learning model with dynamically evolving normality behavior. Therefore, if a detected video frame is an incorrect detection (false positive), then the human observer can mark the video frame as "normal," which will be used in the continuous learning phase.

Subsequent to human observer feedback, the video frames that were marked as normal will be used to continuously train the ISTL model, updating its knowledge of the notion of normality. As shown in Fig. 5, continuous update of the ISTL model is conducted using 1) spatiotemporal information

from the continuously received surveillance video stream and 2) active human observer feedback on detected anomalies.

The continuous learning of the ISTL model is enriched by fuzzy aggregation of video frames, in order to retain stability across iterations of learning. At the detection phase, all the video frames being evaluated are tagged with a fuzzy measure g_λ based on its reconstruction error and grouped into a finite number (n) of sets based on g_λ. Subsequently, in the continuous learning phase, the algorithm will select the k video frame cuboids that contain highest g_λ from each set of fuzzy measures (S) to train the ISTL model. The parameters k and n are defined at initiation based on the duration of video surveillance stream employed for continuous learning. The scene selection for continuous training is defined by (8); $\forall s \in S$, where, $S = \{s_1, s_2, \ldots s_n\}$ and d are the indexes of the selected temporal cuboids that will be included in the continuous training dataset

$$d = \sum_{i=0}^{n} \max_{j=[1,k]} (s_i). \tag{8}$$

The dataset for continuous training iteration is now composed of 1) false positive detection verified by the human observer and 2) temporal cuboids selected across normal behavior using the fuzzy aggregation. This will ensure that the continuous training will update the detection model's capability to capture novel normal behavior while remaining stable for previously known normal behavior. This fuzzy aggregation approach has been successfully demonstrated to maintain stability-plasticity in continuous learning for Internet of Things (IoT) stream mining [38], text mining [17], and video stream mining [24].

Subsequent to the scene selection, the ISTL model will be continuously trained upon the selected representation from input video data, which is the updated expected and acceptable behavior form the surveillance arena. Thenceforth, the updated ISTL model will be re-employed for anomaly detection.

IV. EVALUATION OF THE ISTL APPROACH

The proposed approach, ISTL, is evaluated using three benchmark datasets, CUHK Avenue dataset [21], UCSD Ped 1, and UCSD Ped 2 datasets [20]. With this empirical evaluation, we demonstrate the capability of the ISTL to detect and localize anomalies in near real-time and that the ISTL model performs on par with state-of-the-art anomaly detection methods proposed in the current literature. ISTL was implemented in Python with TensorFlow framework [39] for enhanced capabilities in deep learning and GPU utilization. ISTL was trained on a high-performance computing specification, 36-core CPU 2.3 GHz with 128 GB memory, and dual NVIDIA Quadro of 24GB GPU units. Evaluation of ISTL was conducted on a typical personal computer configuration, a 4-core CPU 2.6 GHz with 24 GB memory, and GPU of NVIDIA GeForce GTX 970M, in order to ensure that the proposed ISTL model can be realistically deployed in an industrial setting.

A. Datasets

The CUHK Avenue dataset [21] was acquired using a stationary video camera with a resolution of 640×360 pixels,

recording street activity at the Chinese University of Hong Kong. This dataset has 16 train video samples that contain normal human behavior and 21 test video samples that contain unusual events and human actions. The normal behaviors are pedestrians on the sidewalk and groups of pedestrians congregating on the sidewalk, whereas the anomalous events are people littering/discarding items, loitering, walking toward the camera, walking on the grass and abandoned objects.

The UCSD pedestrian Dataset [20] was captured by a stationary video camera with a resolution of 238×158 pixels, focusing on two pedestrian walkways. This includes two datasets, Ped 1 and Ped 2, capturing different crowd scenes, ranging from sparse to dense. The normal behaviors of the train video samples contain only scenarios of pedestrians walking on the pathway, whereas the test video samples contain anomalous pedestrian movement patterns such as walking across the sidewalk or walking on the grass, unexpected behavior such as skateboarding, cycling, and vehicular movement. Ped 1 dataset has 34 train video samples and 36 test video samples, whereas Ped 2 dataset has 16 train video samples and 12 test video samples. Both the selected datasets were captured at a frame rate of 26 frames per second (FPS).

B. Experimental Setup

The experimental setup is fourfold.
1) First, the anomaly detection capabilities of the spatiotemporal autoencoder model are evaluated and compared with the state-of-the-art anomaly detection models based on the three benchmark datasets.
2) Second, the anomaly localization capability is evaluated using nonoverlapping cuboids of $16 \times 16 \times T$ pixels. This size is selected for the input cuboid as it is small enough to capture the location of anomalies as well large enough to extract related appearance information, based on the video resolution of selected datasets.
3) Third, we evaluate the continuous learning capability of the ISTL model for UCSD Ped 1 and Ped 2 datasets, adapting a particular scenario as normal which was previously considered as an anomaly.
4) Fourth, we conduct a runtime analysis of our approach demonstrating the real-time processing capabilities of our algorithm.

As the video samples have different dimensionality, we pre-process the inputs by resizing the extracted frames to 224×224 pixels, and normalizing pixel values by scaling between 0 and 1. Based on the frame rate of the selected training data (i.e., 26 FPS), we select the depth of temporal cuboid, $T = 8$ representing an approximate duration of one-third of a second. The selection of T is both dependent on maximizing the motion to be captured within consecutive frames as well minimizing the convergence of the deep learning model due to the large depth of input cuboids. In particular scenarios where the input surveillance data has lower frame rate, it is possible to capture longer motion with low temporal depths.

In this experiment, we trained the learning model using a learning rate of 0.01 and 1500 training epochs. Stochastic gradient descent algorithm is used to optimize the spatiotemporal
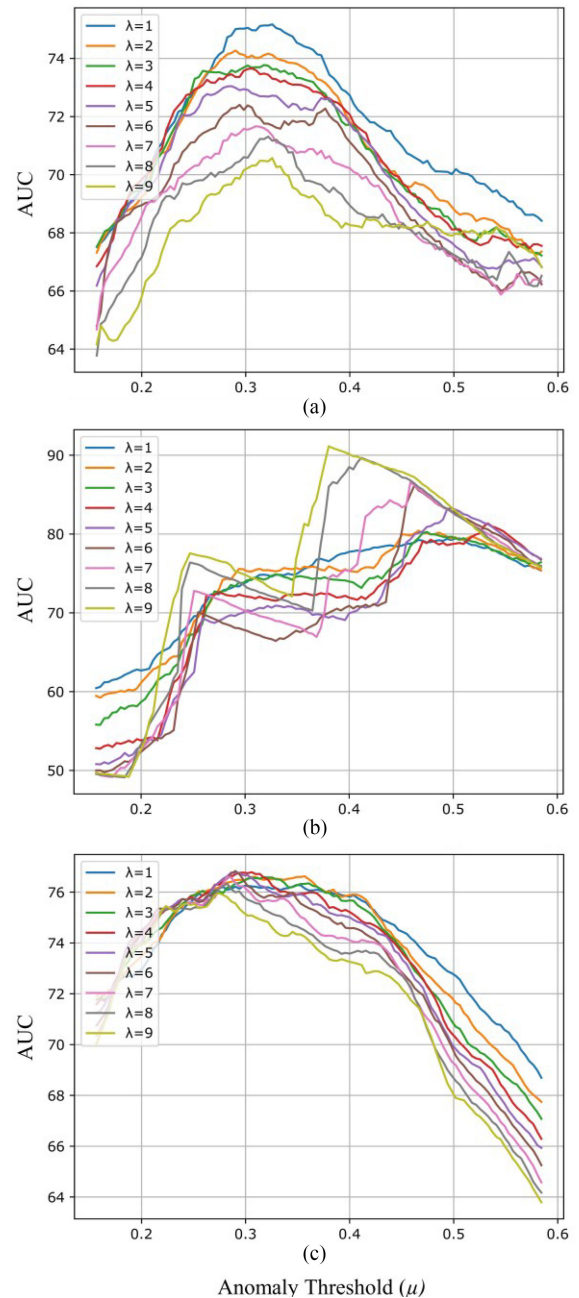

Fig. 6. Evaluation of optimal AUC with respect to μ based on different λ values. (a) UCSD Ped 1, (b) UCSD Ped 2, and (c) CUHK Avenue. Best viewed in color.

autoencoder model and mean squared error is used as the cost function to calculate the reconstruction loss. In order to avoid overfitting of the model, we employed an early stopping regularization technique where the training terminates when the loss has stopped improving. The training was conducted for three continuous iterations by splitting the data set as 60% for the first iteration and 20% each for second and third iterations (as elucidated in Fig. 1). The reconstruction error was used as the fuzzy measure in the active learning phase.

In the anomaly detection and localization phase, the two thresholds are the temporal threshold and the anomaly threshold. We evaluated a range of λ from 1 to 9 in order to select the

TABLE II
SELECTION OF ANOMALY THRESHOLD AND TEMPORAL THRESHOLD

Dataset	Optimal AUC/EER	Anomaly Threshold (μ)	Temporal Threshold (λ)
Ped 1	75.2/29.8	0.33	1
Ped 2	91.1/8.9	0.38	9
Avenue	76.8/29.2	0.29	6

optimal value for each test dataset. The evaluation is presented in Fig. 6. The optimal λ was different for the three datasets; minimum value of 1 (one-third of a second) for Ped 1 dataset, whereas the maximum value of 9 (3 s) for the Ped 2 dataset. This can be justified by the viewpoint of the video samples as Ped 1 dataset has the farthest view making the pedestrian/object movement to be small, whereas the Ped 2 dataset contained a closer view of the pedestrians/objects which made video sample to capture movements a lengthier than the Ped 1 dataset. Avenue dataset captured the optimal anomalies with the λ of 6 (2 s), which similarly can be justified by the camera view and the movement of people captured in the video sample. The optimal accuracies the ISTL model was able to achieve is presented in Table II, with respective λ and μ values.

C. Results—Anomaly Detection

Anomaly detection was evaluated with three state-of-the-art handcrafted feature representation-based approaches and four state-of-the-art deep learning-based approaches. The selected handcrafted feature representation-based methods include, first, abnormal crowd behavior detection using social force (SF) model by Mehran *et al.* [40] which employs a grid of particles is placed over the video frame and the space-time average of optical flow to enforce the SF model. Second, we evaluate MPCCA model (2009) [41] that utilizes space-time Markov random field and video optical flow for anomaly detection. Third, we evaluate MPCCA+SF model (2010) [20], the original work of the UCSD Ped 1 and Ped 2 datasets. The anomaly detection of this approach is based on mixtures of dynamic textures, where the outliers under this model are labeled as anomalies.

The selected deep learning-based approaches for comparison are as follows. First, Conv-AE [25] is a deep convolution feed-forward autoencoder architecture that learns both local features and classifiers as an end-to-end learning framework. Second, streaming restricted boltzmann machine (S-RBM) [31] is an unsupervised probabilistic framework that models the normality and learns feature representations automatically. Third, ConvLSTM-AE [30] is an integrated CNN and ConvLSTM autoencoder to encode spatial and temporal patterns in normal behavior. Fourth, Unmasking-late-fusion [42] is an anomaly detection approach based on unmasking technique. This method employs motion features captured from 3-D gradients and appearance features from pretrained CNN, specifically VGG-f [43].

We compare the results of respective models using frame-level ROC curves, the corresponding area under the curve (AUC) and equal error rate. The comparison is presented in Table III, where the results appear as reported by respective authors. Overall, our method outperforms all the handcrafted approaches

TABLE III
COMPARISON OF AUC AND EER

Model	Ped 1 AUC/EER	Ped 2 AUC/EER	Avenue AUC/EER
SF (2009)	67.5/31.0	55.6/42.0	NA
MPCCA (2009)	66.8/40.0	69.3/30.0	NA
MPCCA + SF (2010)	74.2/32.0	61.3/36.0	NA
Conv-AE (2016)	**81.0/27.9**	90.0/21.7	70.2/25.1
S-RBM (2017)	70.3/35.4	86.4/16.5	78.8/27.2
ConvLSTM-AE (2017)	75.5/NA	88.1/NA	77.0/NA
Unmasking (2017)	68.4/NA	82.2/NA	**80.6/NA**
Ours (ISTL)	75.2/29.8	**91.1/8.9**	76.8/29.2

The bold values in each column represents the best AUC/EER.

whereas we obtain on-par results in comparison to deep learned representation-based methods with respect to Ped 1 and Avenue datasets. For the Ped 2 dataset, our proposed ISTL method outperforms all the compared models including the benchmark of Conv-AE (2016) approach.

D. Results—Anomaly Localization

Qualitative analysis of the localized anomaly patches is presented in Fig. 7. It is shown that anomalies such as cyclists and vehicles on the pathways, pedestrians walking across the pathways, crowd loitering, and pedestrians pushing carts are localized by ISTL in the UCSD Ped 1 dataset. It is important to note that there were false negative detections with respect to skateboarding in Ped 1 dataset [see Fig. 7(a)]. Out of the 12 test videos samples that contained people who skateboard, only ten were detected by the ISTL model. However, in the Ped 2 dataset, all the video samples that contained skateboarding were detected. This can be explained by the camera angle of Ped 1 datasets where its elevation makes it difficult to differentiate between pedestrians and skateboarders by appearance.

In UCSD Ped 2 test samples, bicycles, vehicles, and pedestrians walking in different directions are localized. The main anomaly in the Ped 2 test samples was cyclists, in 11 out of the 12 instances. Anomalies such as an abandoned bag, a person throwing a bag, child playing in the surveillance area, people walking in wrong directions, and people running are localized as anomalies by ISTL in the CUHK avenue dataset.

E. Results—Active Learning

In order to demonstrate the active learning capability of ISTL, we selected cycling on the pedestrian pathway scenarios of UCSD Ped 1 and Ped 2 datasets. Here we defined cycling on pedestrian pathways as a normal behavior, thereby tagged all the anomaly detections from test samples of cyclists as normal. We employed four test samples containing cyclists from each Ped 1 and Ped 2 datasets to continuously train the ISTL model with human observer verification. Subsequent to the training phase, we evaluated the anomalies of the test samples excluding the four samples selected for continuous training. The anomaly detection ratio is presented in Table IV. In Ped 1 dataset evaluation, it was detected that two test samples that had cyclists were anomalous because these were across sidewalk cycle movements.
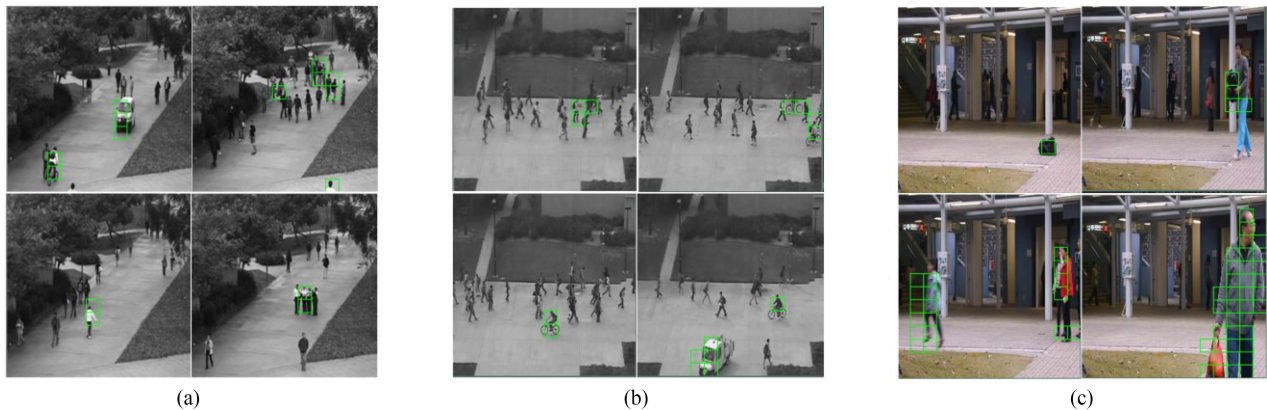
Fig. 7. Localized anomalies. (a) UCSD Ped 1 Dataset, (b) UCSD Ped 2 Dataset, and (c) CUHK Avenue Dataset. Best viewed in color.

TABLE IV
ANOMALY DETECTION FOR CYCLING SCENARIO

Dataset	Prior to active learning	After active learning
UCSD Ped 1	12 / 14	2 / 14
UCSD Ped 2	7 / 7	1 / 7

Values presented as detected anomalies/total test sample.
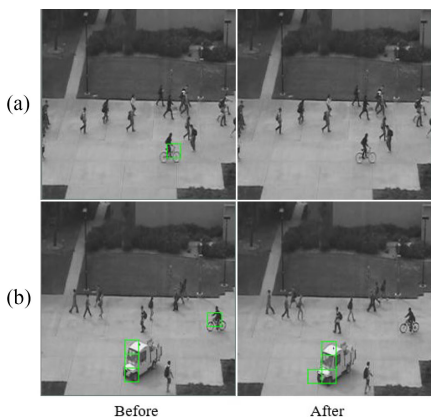

Fig. 8. Evaluation dataset from UCSD Ped 2: (a) person riding a bicycle and (b) person riding a bicycle and a vehicle moving on the pedestrian walk. Best viewed in color.

To further evaluate the utility of the active learning approach, we singled out two particular test scenarios that have been previously detected as anomalous: 1) cyclist only and 2) cyclist and a vehicle moving on the pedestrian pathway (as illustrated in Fig. 8). The evaluation resulted in test video A being detected as normal while test video B being detected as an anomaly. This localization confirms that video B was detected as an anomaly due to the moving vehicle, whereas the cyclist was detected as normal.

F. Results—Run-Time Analysis

We evaluated the real-time video surveillance capability of our anomaly detection approach and the computational overheads for the sequenced process of anomaly detection and localization. Table V presents an overview of the time analysis for our anomaly detection approach in the three datasets

TABLE V
PROCESSING TIME ANALYSIS (SECONDS PER FRAME)

Process	Ped 1	Ped 2	Avenue
Pre-processing	0.0012	0.0012	0.0012
Representation	0.0293	0.0292	0.0290
Detection	0.0019	0.0019	0.0018
Localization	0.0047	0.0049	0.0042
Total	0.0369	0.0371	0.0360
FPS	~27	~27	~28

evaluated. The averaged processing time for anomaly detection and localization is 37 ms. Achieving approximately 27 FPS, ISTL has demonstrated capability for anomaly detection from video surveillance streams in real-time. It should be noted that the difference in processing time for datasets was due to their differences in original resolution as even though frames are resized for anomaly detection, localization is assessed for original frame resolution. For these experiments, ISTL was implemented in series. However, detection and localization can be parallelized, thereby further reducing run time to achieve a higher FPS rate.

V. CONCLUSION

In this article, we proposed a new spatiotemporal anomaly detection approach using deep learning and active learning for real-time video surveillance. This approach addressed the three primary challenges of anomaly detection from surveillance video streams by

1) handling high-dimensional video surveillance data streams in real-time;
2) formulating the anomaly detection as to learn normality;
3) adapting to dynamically evolving normal behavior with fuzzy aggregation and active learning.

The proposed ISTL approach was based on a spatiotemporal autoencoder model consisting of convolution layers that learn spatial regularities and ConvLSTM layers that learn temporal regularities preserving the spatial structure of the video stream. ISTL incorporates a fuzzy aggregation of human observer feedback into a continuous active learning process of unknown/new normalities to address the tightly coupled

dependence on a known normality training dataset. It uses two thresholds, anomaly threshold and temporal threshold, based on the context of the video surveillance feed, to overcome sparse evaluation, which is based solely on reconstruction error.

Results from experiments conducted on three benchmark datasets demonstrate accuracy, robustness, low computational overhead as well as contextual indicators of the proposed approach, confirming its wide applicability in industrial and urban settings. From a practical perspective of video surveillance, ISTL ensures that a human observer is not required to continuously monitor surveillance footage to determine anomalous behavior. Human involvement is only required for verification of the detected anomalies in practical scenarios and refinement of the learning model. As future work, we intend to achieve end-to-end autonomous video surveillance with reduced false negative detection by utilizing hierarchical multistream recurrent self-organizing architecture with transience, in which we process spatial and temporal streams separately so that re-occurring anomalies are retained and long-term anomalies are gradually replaced.

REFERENCES

[1] H. Liu, S. Chen, and N. Kubota, "Intelligent video systems and analytics: A survey," *IEEE Trans. Ind. Informat.*, vol. 9, no. 3, pp. 1222–1233, Aug. 2013.

[2] D. I. Kosmopoulos, A. S. Voulodimos, and A. D. Doulamis, "A system for multicamera task recognition and summarization for structured environments," *IEEE Trans. Ind. Informat.*, vol. 9, no. 1, pp. 161–171, Feb. 2013.

[3] K. Muhammad, R. Hamza, J. Ahmad, J. Lloret, H. H. G. Wang, and S. W. Baik, "Secure surveillance framework for IoT systems using probabilistic image encryption," *IEEE Trans. Ind. Informat.*, vol. 14, no. 8, pp. 3679–3689, Aug. 2018.

[4] L. Zhang, Y. Xia, K. Mao, H. Ma, and Z. Shan, "An effective video summarization framework toward handheld devices," *IEEE Trans. Ind. Electron.*, vol. 62, no. 2, pp. 1309–1316, Feb. 2015.

[5] J. García, A. Gardel, I. Bravo, and J. L. Lázaro, "Multiple view oriented matching algorithm for people reidentification," *IEEE Trans. Ind. Informat.*, vol. 10, no. 3, pp. 1841–1851, Aug. 2014.

[6] Wahyono, A. Filonenko, and K. H. Jo, "Unattended object identification for intelligent surveillance systems using sequence of dual background difference," *IEEE Trans. Ind. Informat.*, vol. 12, no. 6, pp. 2247–2255, Dec. 2016.

[7] M. I. Chacon-Murguia and S. Gonzalez-Duarte, "An adaptive Neural-Fuzzy approach for object detection in dynamic backgrounds for surveillance systems," *IEEE Trans. Ind. Electron.*, vol. 59, no. 8, pp. 3286–3298, Aug. 2012.

[8] A. B. Sargano, P. Angelov, and Z. Habib, "A comprehensive review on handcrafted and learning-based action representation approaches for human activity recognition," *Appl. Sci.*, vol. 7, no. 1, Jan. 2017, Art. no. 110.

[9] G. M. Basavaraj and A. Kusagur, "Vision based surveillance system for detection of human fall," in *Proc. 2nd IEEE Int. Conf. Recent Trends Electron., Inf. Commun. Technol.*, Bangalore, India, May 2017, pp. 1516–1520.

[10] R. Nawaratne, D. Alahakoon, D. De Silva, P. Chhetri, and N. Chilamkurti, "Self-evolving intelligent algorithms for facilitating data interoperability in IoT environments," *Future Gener. Comput. Syst.*, vol. 86, pp. 421–432, Sep. 2018.

[11] B. R. Kiran, D. M. Thomas, and R. Parakkal, "An overview of deep learning based methods for unsupervised and semi-supervised anomaly detection in videos," *J. Imag.*, vol. 4, no. 2, Feb. 2018, Art. no.36.

[12] D. Xu, Y. Yan, E. Ricci, and N. Sebe, "Detecting anomalous events in videos by learning deep representations of appearance and motion," *Comput. Vis. Image Understanding*, vol. 156, pp. 117–127, Mar. 2017.

[13] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Comput. Surv.*, vol. 41, no. 3, pp. 15:1–15:58, Jul. 2009.

[14] A. Gandomi and M. Haider, "Beyond the hype: Big data concepts, methods, and analytics," *Int. J. Inf. Manage.*, vol. 35, no. 2, pp. 137–144, Apr. 2015.

[15] M. M. Najafabadi, F. Villanustre, T. M. Khoshgoftaar, N. Seliya, R. Wald, and E. Muharemagic, "Deep learning applications and challenges in big data analytics," *J. Big Data*, vol. 2, no. 1, Dec. 2015, Art. no. 1.

[16] K. Burbeck and S. Nadjm-Tehrani, "Adaptive real-time anomaly detection with incremental clustering," *Inf. Secur. Tech. Rep.*, vol. 12, no. 1, pp. 56–67, Jan. 2007.

[17] D. De Silva and D. Alahakoon, "Incremental knowledge acquisition and self learning from text," in *Proc. Int. Joint Conf. Neural Netw*, 2010, pp. 1–8.

[18] B. Liu, "Natural intelligence the human factor in AI," in Proc. AI NEXTCon, Jan. 2018. [Online]. Available: https://www.slideshare.net/BillLiu31/natural-intelligence-the-human-factor-in-ai

[19] "How machine learning with TensorFlow enabled mobile proof-of-purchase at Coca-Cola," Google Developers Blog, Sep. 2017.

[20] V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos, "Anomaly detection in crowded scenes," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, San Francisco, CA, USA, Jun. 2010, pp. 1975–1981.

[21] C. Lu, J. Shi, and J. Jia, "Abnormal event detection at 150 FPS in MATLAB," in *Proc. IEEE Int. Conf. Comput. Vis.*, Sydney, NSW, Australia, Dec. 2013, pp. 2720–2727.

[22] D. Gowsikhaa, S. Abirami, and R. Baskaran, "Automated human behavior analysis from surveillance videos: A survey," *Artif. Intell. Rev.*, vol. 42, no. 4, pp. 747–765, Dec. 2014.

[23] S. Ojha and S. Sakhare, "Image processing techniques for object tracking in video surveillance—A survey," in *Proc. Int. Conf. Pervasive Comput.*, 2015, pp. 1–6.

[24] R. Nawaratne, T. Bandaragoda, A. Adikari, D. Alahakoon, D. De Silva, and X. Yu, "Incremental knowledge acquisition and self-learning for autonomous video surveillance," in *Proc. 43rd Annu. Conf. IEEE Ind. Electron. Soc.*, Beijing, China, Oct./Nov. 2017, pp. 4790–4795.

[25] M. Hasan, J. Choi, J. Neumann, A. K. Roy-Chowdhury, and L. S. Davis, "Learning temporal regularity in video sequences," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Las Vegas, NV, USA, Jun. 2016, pp. 733–742.

[26] S. Xie and Y. Guan, "Motion instability based unsupervised online abnormal behaviors detection," *Multimed Tools Appl.*, vol. 75, no. 12, pp. 7423–7444, Jun. 2016.

[27] S. Wu, B. E. Moore, and M. Shah, "Chaotic invariants of Lagrangian particle trajectories for anomaly detection in crowded scenes," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, San Francisco, CA, USA, Jun. 2010, pp. 2054–2060.

[28] B. Zhao, L. Fei-Fei, and E. P. Xing, "Online detection of unusual events in videos via dynamic sparse coding," in *Proc. CVPR*, 2011, pp. 3313–3320.

[29] A. Zaharescu and R. Wildes, "Anomalous behaviour detection using spatiotemporal oriented energies, subset inclusion histogram comparison and event-driven processing," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 563–576.

[30] W. Luo, W. Liu, and S. Gao, "Remembering history with convolutional LSTM for anomaly detection," in *Proc. IEEE Int. Conf. Multimedia Expo*, Hong Kong, China, Jul. 2017, pp. 439–444.

[31] H. Vu, "Deep abnormality detection in video data," in *Proc. Int. Joint Conf. Artif. Intell. Org.*, 2017, pp. 5217–5218.

[32] R. Leyva, V. Sanchez, and C.-T. Li, "Video anomaly detection with compact feature sets for online performance," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3463–3478, Jul. 2017.

[33] M. Sabokrou, M. Fayyaz, M. Fathy, Z. Moayed, and R. Klette, "Deep-anomaly: Fully convolutional neural network for fast anomaly detection in crowded scenes," *Comput. Vision Image Understanding*, vol. 172, pp. 88–97, Jul. 2018.

[34] Y. S. Chong and Y. H. Tay, "Abnormal event detection in videos using spatiotemporal autoencoder," in *Proc. Adv. Neural Netw.*, 2017, pp. 189–196.

[35] P. Baldi, "Autoencoders, unsupervised learning and deep architectures," in *Proc. Int. Conf. Unsupervised Transfer Learn. Workshop*, 2011, pp. 37–50.

[36] M. Matsugu, K. Mori, Y. Mitari, and Y. Kaneda, "Subject independent facial expression recognition with robust face detection using a convolutional neural network," *Neural Netw.*, vol. 16, no. 5, pp. 555–559, Jun. 2003.

[37] S. H. I. Xingjian, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W. Woo, "Convolutional LSTM network: A machine learning approach for precipitation nowcasting," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 802–810.

[38] D. D. Silva, X. Yu, D. Alahakoon, and G. Holmes, "Incremental pattern characterization learning and forecasting for electricity consumption using smart meters," in *Proc. IEEE Int. Symp. Ind. Electron.*, Gdansk, Poland, Jun. 2011, pp. 807–812.

[39] M. Abadi *et al.*, "TensorFlow: A system for large-scale machine learning," in *Proc. 12th USENIX Conf. Operating Syst. Des. Implementation*, 2016, pp. 265–283.

[40] R. Mehran, A. Oyama, and M. Shah, "Abnormal crowd behavior detection using social force model," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Miami, FL, USA, Jun. 2009, pp. 935–942.

[41] J. Kim and K. Grauman, "Observe locally, infer globally: A space-time MRF for detecting abnormal activities with incremental updates," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Miami, FL, USA, Jun. 2009, pp. 2921–2928.

[42] R. Tudor Ionescu, S. Smeureanu, B. Alexe, and M. Popescu, "Unmasking the abnormal events in video," in *Proc. IEEE Int. Conf. Comput. Vis.*, Venice, Italy, Oct. 2017, pp. 2895–2903.

[43] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," in *Proc. Brit. Mach. Vis. Conf.*, 2014.

Daswin De Silva (M'10) received the Ph.D. degree in machine learning and artificial intelligence from Monash University, Melbourne, Vic, Australia, in 2011.

He is currently a Senior Lecturer with the Centre for Data Analytics and Cognition, La Trobe University, Melbourne, Vic, Australia. His research interests include cognitive computing, autonomous learning algorithms, incremental knowledge acquisition, stream mining, social media, and text mining.

Rashmika Nawaratne received the B.S. degree in computer science and engineering from the Department of Computer Science and Engineering, University of Moratuwa, Sri Lanka, in 2014. He is currently working toward the Ph.D. degree in machine learning and artificial intelligence with the Centre for Data Analytics and Cognition at La Trobe University, Melbourne, Vic, Australia.

Prior to commencing his Ph.D., he was with a Software Product Development organization as a Technical Lead. His research interests include self-learning, incremental learning, video analytics, deep learning, and human cognition.

Damminda Alahakoon (M'09) received the B.S. degree in computer science from the Department of Computer Science with the University of Colombo, Sri Lanka, in 1995, and the Ph.D. degree in artificial intelligence from the Department of Computer Science with Monash University, Melbourne, Australia, in 2002.

He is currently Professor and Director of the Research Centre for Data Analytics and Cognition. He has more than 15 years experience as an academic in several Australian Universities as well as more than ten years in the IT and finance industries. His research interests include the areas of data mining, predictive analytics, text analytics, machine learning and business intelligence, and the harnessing of such theories for practical tools and innovative technology for industry.

Xinghuo Yu (M'92–SM'98–F'08) received the B.Eng. and M.Eng. degrees in electrical and electronic engineering from the University of Science and Technology of China, Hefei, China, in 1982 and 1984, respectively, and the Ph.D. degree in control science and engineering from Southeast University, Nanjing, China, in 1988.

He is currently an Associate Deputy Vice-Chancellor and a Distinguished Professor with the Royal Melbourne Institute of Technology, Melbourne, Vic, Australia. His research interests include control systems, complex and intelligent systems, and smart grids.

Dr. Yu is the President of IEEE Industrial Electronics Society for 2018 and 2019. He has served as an Associate Editor of IEEE TRANSACTIONS ON AUTOMATIC CONTROL, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS I: REGULAR PAPERS, IEEE TRANSACTIONS ON INDUSTRIAL ELECTRONICS, and IEEE TRANSACTIONS ON INDUSTRIAL INFORMATICS. He is the recipient of a number of awards and honors for his contributions, including Dr.-Ing. Eugene Mittelmann Achievement Award of IEEE Industrial Electronics Society in 2013, M A Sargent Medal from Engineers Australia and Austalasian AI Distinguished Research Contribution Award from Australian Computer Society, in 2018. He was named a Highly Cited Researcher by Clarivate Analytics/Thomson Reuters consecutively in 2015–2018.