Unusual Human Behavior Detection System in Real-Time Video Systems

Yanbin Bu, Ting Chen*, Hongxiu Duan, Mei Liu, Yandan Xue
School of Media Technology, Communication University of China Nanjing, Nanjing 211172, China

***Abstract*—Abnormal behavior detection, in terms of importance, has become a necessity in real-time visual systems. The main problem is the ambiguity in the difference between the characteristics of abnormal and normal behavior, which its definition is usually different according to the previous context of images. In this research, three approaches are used. In the first approach, a standard Convolutional Automatic Encoder (CAE) is used. After evaluation, it was found that the standard CAE problem is that it does not take into account the temporal aspect of the image frames sequence. The second method involves automatic encoding to learn the dataset's spatio-temporal structures. In the third approach, the complex LSTM cells are used for further improvement. The outcomes of the test display that the proposed methods have better performance compared to many of the previous conventional methods, and their efficiency in identifying abnormal behavior is very competitive compared to previous methods.**

Keywords*—Anomaly detection; video sequence; standard Convolutional Automatic Encoder (CAE); spatio-temporal structures; LSTM

I. Introduction

Over the past decade, real-time video analytics has grown rapidly and made tremendous progress. The primary goal of video analysis is to identify possible occurrences with minimal (or no) human intervention. Studies in the popular field of video control involve recognizing human operations as well as classifying these operations as usual and unusual or suspicious operations. The main role in this area is to identify abnormal events in the videos using a monitoring system (which is fully automatic, manual, and semi-automatic). The manual monitoring system is completely human-dependent. Manual activity is required to analyze human behavior or to distinguish between abnormal and natural behaviors. The semi-automatic system requires less human interposition, while the fully automated surveillance system is the intelligent video surveillance system that is fully automatic and does not require human intervention to make decisions.

According to the current observations in the market [1], the public and private sectors invest a lot of money to protect the offices, buildings, centers, houses, infrastructure, etc.; these trends in the coming years will improve the automated security industry. As terrorist activities are on the rise today, it is important to identify the suspicious or abnormal operations that could affect the usual human operations. The unusual events are the disorders or behavioral deviations of an object (relative to the usual behavior of that object) that include placing the object in an abnormal location, unusual movement pattern (such as moving in the mistake direction, unusual rotation, violence or conflict between people or different movements as opposed to general movements such as walking all people but crawling some people) or any the abnormal event. Each event can be normalized in one scenario and abnormal in another [2].

The recognition of unusual events can be done in two techniques: The first is to train the system with usual events and unusual events and then to identify future occurrences using the previous information. The second method is to follow the dominant ownership according to which the dominant behavior of the individual (behavior that occurs frequently) is considered normal behavior, and the behavior that occurs less often is considered unusual and unusual. An anomaly is detected by taking and analyzing the motion and physical signs of the objects in the video [3]. The method of detecting motion anomaly includes the speed, direction, location and path of the moving object. The method of detecting outward anomalies includes the condition of the object, the identification or color of the object, and so on [4].

In this article, various systems are presented that detect anomalies in video frames. In this research, three approaches are used, and the results of each are presented. In each approach, by adding more aspects, it is possible to improve the previous approach. Therefore, the main contributions of the current article can be stated as follows:

- Improving the performance of this system by providing different methods and considering the temporal and spatial sequence of the frames.
- Increasing the speed of this system to detect abnormal behaviors
- Designing the appropriate methods to identify the various abnormal behaviors according to the selected dataset.

The remaining article: Related research in this field is described in the second section. The suggested methods is explained in the third section. The dataset's elements as well as the implementation results are provided in the fourth part. This paper's conclusions and future research directions are also stated in the fifth part.

II. Related Works

In this part, previous work or existing research background in the area of automatic detection of unusual human activities is discussed. Various frameworks can be used to recognize abnormal behaviors in surveillance video (without human interposition) [5]. Researchists have used different methods depending on the application or events under study. In previous

*Corresponding Author.

research on recognizing abnormal events, Young et al. [6] suggested a new approach to detecting unusual behaviors and identifying dominant behaviors. The dominant set theory proposed by Alvar et al. [7] is used to recognize the abnormal behavior of the object in the image frames sequence. In this regard, Wang et al. [8] suggested using the covariance matrix as a feature descriptor. The SVM classification of an online nonlinear is used to classify the usual and unusual events.

Chung et al. [9] have presented a review article that describes how to solve the problem of showing abnormal event videos. The concept of the conditioned finite Boltzmann machine and the independent component analysis has been used to extract better features (as a common easy method). It was discussed to learn fully learned feature representations and the concept of in-depth feature analysis. In identifying the activity, the most important and vital task is understanding the behavior that Jiang et al. [10] tested and evaluated in a review article. In this work, the author describes (in detail) the characteristics of the behaviors in the videos to detect the anomaly. Different parts of the body are used to recognize human gestures and emotions. Chen et al. [11] introduced a common time filter approach in which the head and the other parts of the body are used to extract features as well as to analyze human behavior. The research in this field has faced the problem of using related examples in multi-factor scenarios. Zhou et al. [12] have proposed the concept of feature labeling with multi-level correlation in videos to identify the different events.

A multi-feature-based method proposed by Hong et al. is used to detect and track different objects [13]. Daphner and Garcia [14] proposed a method that uses pixel-based descriptors to detect very small objects in the image. Also, Ning et al. [15] presented a common recording approach and a smart contour segmentation method for object tracking. Zhang et al. [16] presented the heavy obstruction in object tracking using the outward model. The spatio-temporal model is used to track object videos with obstruction. Several methods have been proposed to detect abnormal behavior. Depending on whether the sample videos require initial determination or training before detecting any unusual operation [16], the supervised approach, the semi-supervised approach, and the unsupervised approach are the three categories into which these techniques fall.

In the supervised approaches, the anomaly recognization input samples are labeled the usual and unusual [17]. The technique is prepared for activities with predetermined features, and path, movement, speed, or appearance is utilized as indicators for classifying them into normal and abnormal categories. The second method is a semi-supervised approach that requires only natural information to train the system. The following categories are used for categorizing this technique: model-based classification and rule-based classification. In the rule-based approach, the rules are set, which help classify the sample into two categories: normal and abnormal. Samples that comply with the rules are classified as normal behaviors, and samples that do not comply with the rules are classified as abnormal. Online dictionary update and flexible encryption [9, 18, 19] are two techniques primarily used in the rule-based approach. The third strategy is unsupervised, which does not require both usual and unusual cases as training data. In these approaches, the classification is based on the hypotheses that state that abnormal behaviors occur less frequently (compared to normal behaviors).

A. Limitaions of Related Works and Solutions

During the past two decades, the recognization and the tracking of the humans in the consecutive video frames, representing and analyzing their activities, and finally identifying their intrusive behavior has been one of the most challenging topics in the field of machine vision, and the attention of the research groups has attracted many reputable universities. On the other hand, the detection of abnormal behavior in video frames faces many limitations. According to the background of the presented research, some of these limitations (that the related works are faced) is briefly listed. The presented method in [28,38], which focuses on the detected paths of objects, assigns the normal labels and the abnormal labels based on which conventional path is ahead. These methods noticeably lose their effectiveness when there is an obstruction or when there is a change in the brightness of the images, and also when there are crowded scenes in the images, these methods have a high computational complexity. Therefore, researchers have proposed the methods that use low-level features such as hinges and gradients to learn the spatial-temporal dimensions and relationships in such features. On the other hand, in some researches such as [12], the one-class SVM classifier was used in the upper layer, which is also a challenge because the detection of the anomalies in video frames that related data have not class label, is not possible. Therefore, methods that are compatible with the non-labeling data class should be considered. Also, the most important challenge in some of the works done in this field such as [43,36] is that the proposed methods are applied to specific data and video frames, which have limitations. There are some of them, the most important of which is not covering a large number of the abnormalities. For example, the UMN dataset [32] and Hockey Fight [33] only include the fight anomaly. For this purpose, it is necessary to consider a dataset that includes a wider range of anomalies. In addition, some works such as [5,35] work on the basis of extracting features from the detected paths of the objects that do not consider the aspect of temporal and spatial sequence of the video frames. This leads to not identifying the certain abnormalities. Therefore, in order to further improve the anomaly detection systems, it is felt necessary to use the aspect of temporal and spatial sequence of the frames. It makes the frames share their learning between the adjacent frames and then reduce the processing cost.

According to the mentioned limitations, the points that are considered to solve these limitations in this article are as follows. The first point is that due to the fact that the data does not have a class label, in this research, the auto-encoders were used to encode and decode the video frames to overcome this limitation. The second point is that in the selection of the dataset, the current article has considered a dataset that includes a wider range of anomalies. This dataset includes thee movement abnormalities of cyclists, skaters, motorbikes, small carts, people in wheelchairs, etc. The third point is that the different methods have been tried to provide more

improvement, and also the aspect of temporal and spatial sequence has been considered.

III. Proposed Methods

The methods used in this study are according to the point that when an unusual event happens, the newest video frames differ from the old frames. An end-to-end model is trained with a feature descriptor as well as a decoder-encoder that trains the frame input volume patterns in a manner that is inspired by [20]. This model is trained with the input video, so these video volumes consist only of frames with normal behavior. This work aims to reduce renovation error, which is the difference between the input video volume. After proper model training, the usual video is awaited to have a low renovation error. However, the video frames are expected to consist of frames with abnormal behaviors also high renovation errors. By limiting the error generated by each input value, the system can recognize when an unusual event happens [21, 22]. In general, the presented method includes three main steps: pre-processing, feature learning (which, in this study, three learning approaches is used), and regularity score.

A. Pre-processing

At this stage, the conversion of raw data into balanced and acceptable input for the model is done. To assure that the input frames are the same scale, each is extracted from the input video and resized to 100×100. Next, the pixel values are scaled between zero and one; for normalization, each frame is subtracted from its global average image. The average image is computed in the training dataset by averaging the pixel values in each position in each frame. The photographs are then made into grayscale versions to make them smaller. In order to have a single mean and variance, the photos are then standardized [23].

The input of the model in some of the approaches used in this research is the volume of the video, in which each volume contains 10 continuous frames with different steps. A lot of training data is needed because this method has a lot of parameters. Therefore, to increment the amount of the training dataset, the data in the time dimension is reinforced. To produce these volumes, the frames with step-1, step-2 and step-3 is connected. For instance, the first sequence of step-1 consists of frames (1,…,10), while the sequence of the first sequence of step-2 contains a numbered frame (1, 3, 5, 7, 9, 11, 13, 15, 17, 19) and the first sequence of step-3 includes frames (1, 4, 7, 10, 13, 16, 19, 22, 25,28). The input is now ready to train the model [23].

B. Feature Learning

As mentioned earlier, three approaches are used to creating regular patterns in the training videos. In the first approach, a standard convolutional automatic encoder (CAE) is used. After using CAE, it became clear that the problem with standard CAE is that it does not take into account the temporal aspect of the image sequence. Thus, it is not easy to identify certain abnormalities, such as a person moving faster than average. Therefore, in the second approach, an automatic encoder is used to learn spatio-temporal structures in the dataset. That is, instead of considering only one image at a time, n images are considered simultaneously. In the third approach, complex LSTM cells are used for further improvement. LSTMs can be used to predict the next video frames. Below, each of the approaches and their details is described:

1) Standard convolution automatic encoder: According to its name, the automatic encoder contains two stages: encoding and decoding. By adjusting the numeral of encoder output modules to be less than the input, an automatic encoder has been employed for the first time to reduce dimensions. This model is trained using error replication in an unsupervised method and minimizing the renovation error of decoding outcomes from the original inputs. An automatic encoder can extract more advantageous information by choosing the nonlinear activation function over traditional linear conversion techniques like PCA. Here, these automatic encoders in the unsupervised method are used to detect the anomalies because a supervised learning method suffers from an imbalance [24].

However, automatic encoders are great for this status because these encoders can be trained on usual components and do not require marginal data. After the training, a feature view is provided for a section and compares the output of the automatic encoder with the input. If there is more difference, the more likely it is that the input contains anomalies. As mentioned, the automatic encoders consist of two sections: 1) an encoder that encodes the input data using a reduced representation and 2) a decoder that attempts to renovate the original input data from the reduced representation. The network is subject to restrictions that force the automatic encoder to learn a concise representation of the training dataset. It does this in an unsupervised manner and is, therefore, the most appropriate case for abnormalities [25].

Here, the used network structure is defined. The encoder includes two layers of convolution and two layers of MaxPooling. The decoder and encoder are connected by a fully connected layer. The bottleneck larger can be reconstructed the more information. The decoder includes two upsampling layers and two deconvolutions' layers. Fig. 1 shows the presented network structure. So, with using CAE, it becomes clear that the problem with standard CAE is that it does not take into account the temporal aspect of the image sequence. Thus, it is not easy to identify certain abnormalities, such as a person moving faster than average. Therefore, in the second approach, an automated encoder is used that can learn spatio-temporal structures in the dataset. That is, instead of considering just one image at a time, n images are considered simultaneously, which is explained in the next section.

2) Spatio-temporal stacked frame encoder: The proposed architecture presented in this approach includes two sections: 1) the automatic spatial encoder for learning the spatial structures of each frame and 2) the temporal encoder-decoder for learning the temporal samples of encoded spatial structures. The spatial encoder and decoder, as seen in Fig. 2, have two layers of convolution and deconvolution, respectively. Convolution layers are renowned for their superior object detection performance. Convolution in a convolution network primarily extracts the necessary features from the input image. By understanding image attributes,

convolution preserves the spatial link among pixels (employing tiny input data squares) [20].

Convolution operations are point multiplications between the local input areas and the filters mathematically. Suppose the several square input layers $n \times n$ be available, followed by the convolution layer. If the $m \times m$ filter is used, then the output of the convolution layer will be $(n - m + 1) \times (n - m + 1)$. During training, a convolutional network discovers the filters' values independently. However, before training, the parameters like the numeral of filters, the size of the filters, and the numeral of layers should be defined. More filters enable us to extract more image features, and the resulting network is better at spotting patterns in previously unseen images. A balance should be striked by not altering the number of very large filters because more filters need more computation time and use memory faster [26].

It is presumed that all inputs (and outputs) in a traditional feed-forward neural network are independent. However, learning the temporal dependencies between the inputs (in sequence tasks) is important. A word prediction model, for instance, should be able to gather data from previous inputs. The RNN functions identically like a feed-forward network, except that the input vector and the complete input history impact the output vector values [20].

Theoretically, RNNs could use arbitrary long sequences of information, and however, in reality, RNNs are constrained to a few steps because slopes have disappeared. On the other hand, as mentioned earlier, a problem with the standard CAE is that it does not take into account the temporal aspect of the image sequence. Thus, identifying specific abnormalities, such as a person moving faster than average, is not easily detectable [20].Therefore, in this approach, an automatic encoder is described that can also learn the spatio-temporal structures in a dataset. In this approach, instead of considering just one image at a time, n images are simultaneously considered. The standard CAE considers input as [packet size, 1, width, height], and the spatio-temporal encoder considers input as [packet size, size, width, height]. In the third approach, the complex LSTM cells are used for further recovery. LSTMs can be used to predict the next frames of a video, and their details are given in the next section.

This architecture receives a trail of length T as an input sample as well as generates a renovation of the input sample trail. Each layer's outcome size is indicated by the numbers on the right. Every time, the location encoder collects one frame as input. It processes 10 frames, delivers the attributes encoded in 10 frames, and gives the encoder time to complete the encoding. Decoders mimic encoders in the reverse direction to reconstruct the volume of the video.

3) Spatio-temporal auto-encoder with convolutional LSTMs: In this section, the short memory model is used and add it to the third approach for further improvement. In other words, the second approach is developed by using LSTM. As mentioned in the second approach, the spatio-temporal encoder and decoder both include two layers of convolution and deconvolution. In contrast, the temporal encoder (in the third technique) adds three-layer long short-term memory (LSTM) convolution. The LSTM model is popular for sequence learning and time series modeling and has demonstrated its performance in applications like speech translation and handwriting identification. Convolution layers are known for their high performance in object recognition [23]. The general architecture of the third approach's proposed method is depicted in Fig. 3.

In the previous section, a brief description of RNN was given. In this approach, as stated, a type of RNN is used: the forgetfulness gate, which is a return gate in the long short-term memory model (LSTM). With this suggested structure, LSTMs are prevented from dissipating or exploding post-propagation errors, allowing them to act on lengthy trails and be combined to gain higher-level information. Equations 1 to 6 and Fig. 4 provide the tabloid formulation of a common LSTM [23]:

$$f_t = \sigma\big(W_f \otimes [h_{t-1}, x_t] + b_f\big) \quad (1)$$

$$i_t = \sigma(W_i \otimes [h_{t-1}, x_t] + b_i) \quad (2)$$

$$\grave{C}_t = tanh(W_C \otimes [h_{t-1}, x_t] + b_C) \quad (3)$$

$$C_t = f_t \otimes C_{t-1} + i_t \otimes \grave{C}_t \quad (4)$$

$$o_t = \sigma(W_o \otimes [h_{t-1}, x_t] + b_o) \quad (5)$$

$$h_t = o_t \otimes tanh(C_t) \quad (6)$$

Equation 1 displays the layer of forgetfulness. Equations 2 and 3 are where new data is subjoined, and Equation 4 merges new and old data, while Equations 5 and 6 are moved from the current LSTM unit and apply what has been previously trained in the upcoming time step. The variable h_t indicates the latent state, x_t indicates the input sample, also C_t indicates the cellular state at time t. b is a bias vector, and W is a teachable matrix, and the symbol $\otimes$ represents the product of Hadamard [23].

The convolutional long short-term memory model (ConvLSTM), a type of LSTM architecture, was first presented by Shi et al. [27]And more recently used by Patrakin et al., which presented in [28]and is used to predict the video frames. In comparison to conventional fully connected LSTM, ConvLSTM replaces its matrix operation with convolution. ConvLSTM requires less weight and provides a map with better spatial features by employing convolution for hidden-to-hidden and input-to-hidden connections. Equations 7 to 12 can be used to summarize the ConvLSTM unit's formulation [23].

$$f_t = \sigma\big(W_f * [h_{t-1}, x_t, C_{t-1}] + b_f\big) \quad (7)$$

$$i_t = \sigma(W_i * [h_{t-1}, x_t, C_{t-1}] + b_i) \quad (8)$$

$$\grave{C}_t = tanh(W_C * [h_{t-1}, x_t] + b_C) \quad (9)$$

$$C_t = f_t \otimes C_{t-1} + i_t \otimes \grave{C}_t \quad (10)$$

$$o_t = \sigma(W_o * [h_{t-1}, x_t, C_{t-1}] + b_o) \quad (11)$$

$$h_t = o_t \otimes tanh(C_t) \quad (12)$$

While these equations are similar to Equations 1 to 6, their input is in the form of an image. At the same time, the weight set for each connection is replaced by convolution filters

(symbol $*$ indicates a torsional action). This allows ConvLSTM to work with better images than FC-LSTM because it can propagate the spatial features (per unit time) through any ConvLSTM mode. Note that this convolution type also adds the optional hole connections to allow a unit to receive the previous information better. So, the previous model is developed by using complex LSTM cells. This proves that ConvLSTM is more efficient in video processing, and ConvLSTM can also be used to predict the next video frames [23]. In this study, 10 input frames are stacked in a cube. These frames placed on the cube are processed by 2 layers of convolution (encoder). Then, these are given to a temporal encoder/decoder consisting of 3 layers of LSTM convolution and 2 layers of deconvolution and the output frames are reconstructed. When initializing the model, the initial state vector for LSTMs should be created.

C. Regularity Score

After model training, the input of experimental data into the trained model can be used to analyze the efficiency of presented models and examine whether these models can reduce the detection of false abnormal behaviors. Also, it examines whether these models can detect abnormal events correctly or not. For better comparison, the regularity score is calculated using the same procedure for all image frames; the only variation is in the model that was learned. The Euclidean distance between the input frame and the renovated frame is calculated using the renovation error of all values of pixels in the frame t of the video trail [23]:

$$\mathrm{e(t)} = \| x(t) - fw\big(x(t)\big) \|_2 \qquad (13)$$

where, fW is the weight training by the spatio-temporal model. Then, the anomaly score $s_a(t)$ is calculated by scaling between zero and one. The regularity score $s_r(t)$ can thus be readily calculated by subtracting the anomaly score by one [23]:

$$s_a(t) = \frac{e(t) - e(t)_{min}}{e(t)_{max}} \qquad (14)$$

$$s_r(t) = 1 - s_a(t) \qquad (15)$$

D. Anomalies Detection

1) Thresholding: It's easy to tell whether a video frame is common or exceptional. Each frame's renovation error indicates whether it may be considered an abnormal frame. This threshold specifies how much of the sensitivity of behavior recognition system. Setting a low threshold, for example, makes the system more sensitive to scene events, which causes more alerts to be generated. Setting various error thresholds allows us to obtain the false positive and true positive values, which are then used to compute the area under the receiver operating characteristic (ROC) curve. Additionally, when the false positive rate is the same as the false negative rate, the equal error rate (EER) is gained [29].

2) Events count: As described in [20], the Persistence1D algorithm is utilized to simultaneously group the local minimums with a fixed temporal frame of 50 frames to reduce noise and meaningless minimums in the regularity score. Therefore, it is presumed the local minimum of 50 frames refers to the same unusual event. This is an advisable temporal window size because an unusual event must take at least 2-3s to make sense (videos are recorded at a speed of 24-25 frames per second).

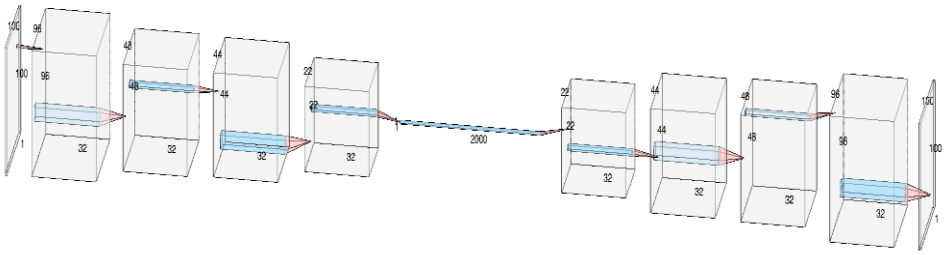

Fig. 1. Proposed network structure.

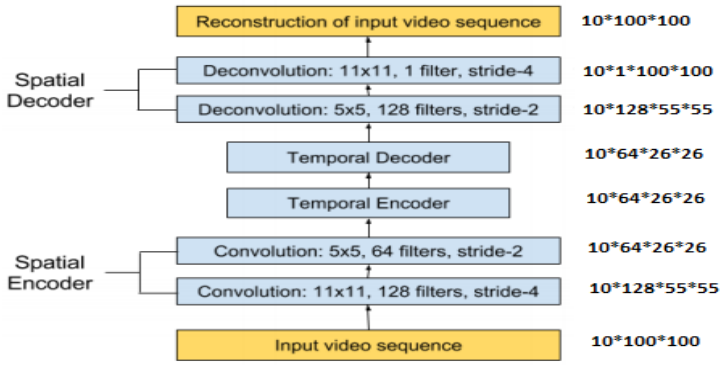

Fig. 2. Proposed network architecture for the second approach.

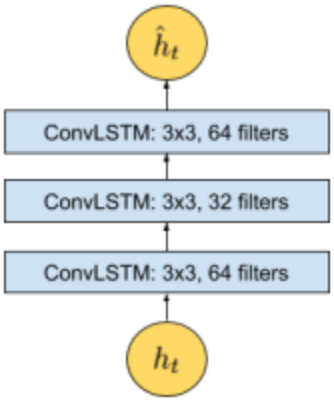

Fig. 3. Magnified architecture at Time t for the third approach where t is the input sample at this Time point. There are three layers of convlstm in the temporal encoder-decoder pattern.

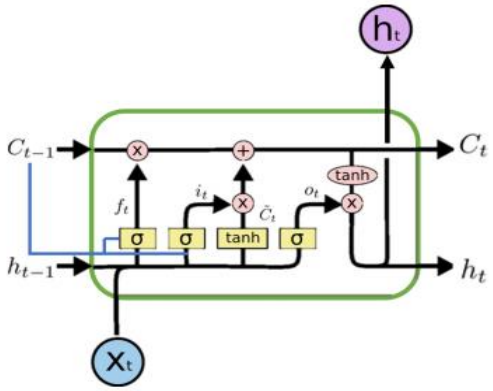

Fig. 4. A generic LSTM's architectural structure. The blue line displays an optional hole structure that allows the inner state to see the state of the prior cell C_{t-1} for better decision-making (best seen in color).

IV. Dataset and Implementation Results

In this part, the implementation's details and its outcomes is described. The presented methods in this study have been compared with previous conventional methods on the UCSD dataset, which will be described below. This comparison has been evaluated and concluded using ROC curve criteria and EER rate. A scatter plot of sensitivity for a binary classifier model with a variable threshold is the ROC curve. In order to specify the abnormal frames, the levels of frame pixels are used. These two measurement values are defined below:

- Measurement at the frame level: If a single pixel of each frame is considered unusual, then this frame is considered abnormal.
- Pixel-level measurement: if the algorithm detects that at least 40% of the correct background pixels are covered by unusual pixels, and then the corresponding frame is considered as abnormal.

A. Dataset

In order to train the model and evaluate the presented method, the UCSD dataset is used, which has been used in almost all the articles presented in this field. Two separate open spaces were used to collect the two subsets of this dataset, Peds1 and Peds2, respectively. A fixed camera recorded both subsets at 10 frames per second at a resolution of 158×234 and 240×360. The right background files in it allow evaluating the levels of the frame and pixel. Thirty-six videos were used for testing in the Peds1 subset, 34 videos for training in the Peds2 subset, 16 videos for testing and 12 videos for training. Fig. 5 displays an instance of frames that existed in this dataset.

In the proposed method, a pre-processing step is presented in which the conversion of raw data into a smooth as well as acceptable input for the model is performed. Every frame is taken out of the raw video and resized to 100×100 to verify that the input frames have the same scale. Next, the pixels' value is scaled between zero and one; for normalization, each frame is subtracted from its global average image. The average image is computed in the training dataset by averaging the value of pixels at each position in each frame. To lower the size, the images are then changed to grayscale versions. In order to have a single mean and variance, the output images are then normalized. Then, the input of the model in some of the approaches used in this research is the volume of the video, in which each volume contains 10 successive frames with different steps. Large amounts of training data are needed because this model has a lot of parameters. Therefore, to augment the length of the training dataset, the data in the time dimension is reinforced. To create this volume, the frames with step-1, step-2 and step-3 is connected [23].

B. Evaluation Criteria

The presented methods in this research have been evaluated with some of the common methods that have been presented so far [30-38]. In order to evaluate the methods presented in the UCSD dataset, two criteria have been used to evaluate the accuracy of detecting the unusual behaviors: the pixel-level paragon also the frame-level paragon. The frame level paragon centralizes only on changes that predict which frame contains the unusual behavior without specifying where it occurs. A frame is regarded as abnormal by the frame-level paragon if it has at least one abnormal result and is not sensitive to the frame's location of the abnormal behavior. Also, the pixel level paragon is a measure that determines the temporal-spatial position of the frame. As mentioned, if at least 40% of the background pixels are correctly covered by pixels that the algorithm recognizes as unusual behavior, it will recognize that frame as unusual.

Then, by calculating the TP and FP rates, the ROC criterion can be obtained to analyze the algorithm's efficiency.

1) Implementation Results: As stated in the prior section, the images are selected at 100×100. Since none of the training videos in the Peds1 and Peds2 sets contains anomalies, half of the testing videos related to the Peds1 and Peds2 are randomly assigned for use in the training model. The remaining videos are used to test the samples. Each of the Peds1 and Peds2 are taught separately. 140248 normal samples and 35215 abnormal samples were extracted from the Peds1 set, and 63579 normal and 20638 abnormal samples were extracted from the Peds2 set.

Because abnormal samples are substantially fewer in number than normal samples, there may be an imbalance problem in the class. To solve this problem by re-sampling, the number of the usual and unusual instances was tried to be in equilibrium. The presented method in this research is supported by a software interface called Keras [39], which supports the neural networks and convolutional using the Theano [40, 41] and TensorFlow software libraries [42]. This software interface is written in Python and can be run on CPU and GPU.

GHz Intel (R) Core (TM) i7 CPU and 8G RAM were used in the proposed approach. The convolutional network is implemented in GPU, and the graphic card used in this method is NVIDIA GEFORCE 840M. The results of the method are presented, as well as previous research, on Peds1 and Peds2 of the UCSD dataset in the following sections. The results obtained from the other methods are adapted from the relevant references in which these methods are introduced.

On the other hand, the system is trained by minimizing input volume renovation errors. The mini-batch with size 64 is used and every training session lasts up to 50 epochs or until the data validation loss stops after 10 successive epochs. The spatial encoder and decoder's activation function is chosen to be the hyperbolic tangent. Despite its ability to adjust, the re-modified linear unit (ReLU) is not used to verify the polarity of the encoding and decoding function since the activation values from ReLU are not very high. An example of the output and abnormal behaviors detected in the UCSD dataset is depicted in Fig. 6. This figure shows the output of identifying abnormal behaviors for the various methods presented in this study. Below, the results related to the error as well as the accuracy of the suggested methods are displayed. The methods presented in this research have been evaluated with some common methods presented so far [30-38, 43]The results related to the presented methods as well as previous studies, are displayed in Tables I and II of the UCSD dataset in Peds1 and Peds2, respectively.

Fig. 5. Examples of images in the UCSD dataset: The first row is Peds1, and the second is Peds2.

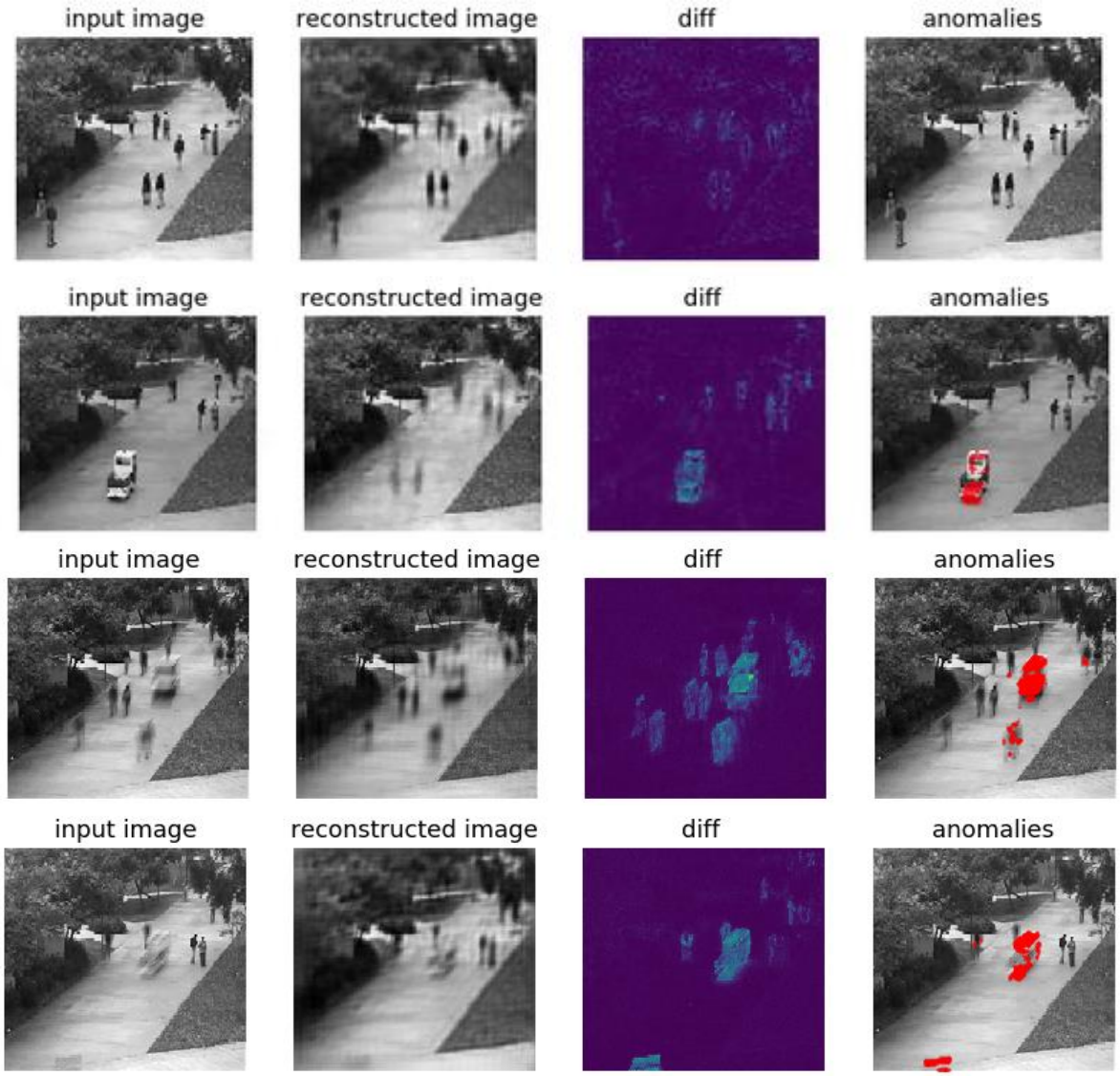

Fig. 6. Output of identifying abnormal behaviors for different methods.

TABLE I. ERR RATE IN THE PEDS1 SUBSET OF THE UCSD DATASET

Name of the author of the article	ERR rate at the pixel level	ERR rate at the frame level
Cheng et al. [31]	38.8	19.9
Cong et al. [32]	51.2	23
Adam [30]	38.9	23.6
Kim [35]	39.6	19.6
Kaltsa [34]	27	21.1
The first proposed method	49.7	33.1
The second proposed method	28.5	21.9
The third proposed method	27.5	21.1

TABLE II. ERR RATES IN THE PEDS2 SUBSET OF THE UCSD DATASET

Name of the author of the article	ERR rate at the pixel level	ERR rate at the frame level
Adam [30]	43.8	22.4
Kim [35]	31.1	22.4
Kaltsa [34]	26.9	25.1
The first proposed method	48.8	35.4
The second proposed method	27.9	20.2
The third proposed method	26.8	19.1

The first row is the output for standard automatic convolutional without the dense layer; the second row is the output for standard automatic convolutional with the dense layer; the third row is the output of temporal-spatial automatic convolutional; fourth row is the output for temporal-spatial automatic convolutional with LSTM.

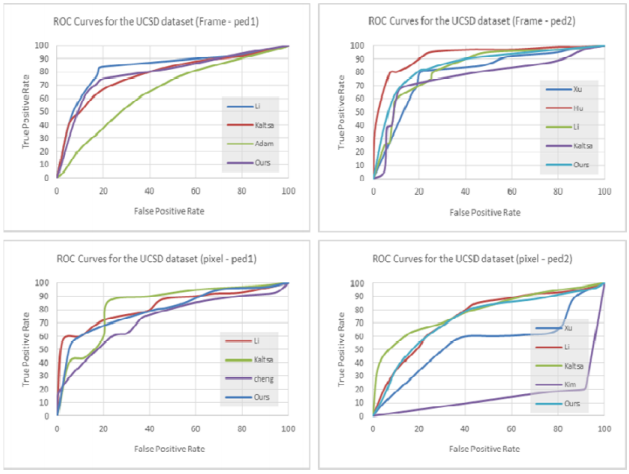

Fig. 7. UCSD dataset ROC curves at the pixel and frame levels.

The obtained results from the other methods are adapted from the relevant references in which these methods are introduced. The results of detecting abnormal behaviors are given in the figures presented above. The ROC curves in this dataset are reported in Fig. 7 for best presented method versus other methods. According to the image below, it can be seen that the presented methods have shown competitive performance in comparison with the previous methods presented in this field. In particular, the results obtained at the frame level in the proposed methods are very similar to the results of the best approaches introduced in this field. Compared to other methods, the suggested methods have a competitive efficiency at the pixel level. In general, it can be said that the methods presented in this study have a very competitive efficiency in the UCSD dataset with the results of other methods.

V. Conclusions and Suggestions

This study addresses the challenging issue of video anomaly recognition using deep learning. In this study, three approaches are presented. The anomaly detection is formulated as a distance detection problem in the spatio-temporal sequence. The best approach to solve this problem is to compound the ConvLSTM and the spatial feature extractor. In this approach, which works best, the ConvLSTM layer retains the benefits of FC-LSTM and is also appropriate for spatio-temporal data due to its inherent convolution structure. By using a spatial and temporal convolution feature extractor in the encoder-decoder structure, a trainable model has developed for detecting video anomalies. The presented models have the benefit of being semi-supervised; all that is required is a lengthy movie with just typical events in a still view. Notwithstanding the models' ability to detect unusual events and their power against existing noise, these methods might produce more erroneous alarms than other techniques depending on how complicated the scene's activities are.

In the future, researchers can examine ways to improve the results of video anomaly recognition with active training; another direction is to consider using human feedback to update the trained model for better recognition and fewer false alarms. One solution is to add a monitored module to this system that will only work on the video segments that have been filtered by using the way which have described. After gathering enough video data, it trains a differential model for classifying the anomalies.

Acknowledgment

This work was supported by the Special Project of Philosophy and Social Sciences Research Ideological and Political Work of Jiangsu Province Higher Education Institutions(2022SJSZ0219), and Special Projects of Jiangsu Higher Education Association (2021JDKT065) (2022JDKT128), and Project of the 2022 National Association for Basic Computer Education in Higher Education Institutions: (2022-AFCEC-410)

References

[1] Yunpeng Chang, Zhigang Tu, Wei Xie, and Junsong Yuan. "Clustering driven deep autoencoder for video anomaly detection". In European Conference on Computer Vision, pages 329–345. Springer, 2020.

[2] A.B. Nassif, M.A. Talib, Q. Nasir, F.M. Dakalbab, "Machine learning for anomaly detection: A systematic review", Ieee Access, 9, pp. 78658-78700, 2021.

[3] A. Azam, K. Singh, "Road Accident Prevention Using Alcohol Detector and Accelerometer Module", EasyChair, 2021.

[4] S. Deepak, P.M. Ameer, "Automated categorization of brain tumor from mri using cnn features and svm", Journal of Ambient Intelligence and Humanized Computing, 12, pp. 8357-8369, 2021.

[5] S.-R. Ke, H.L.U. Thuc, Y.-J. Lee, J.-N. Hwang, J.-H. Yoo, K.-H. Choi, "A review on video-based human activity recognition", Computers, 2, pp. 88-131, 2013.

[6] M. Javan Roshtkhari, M.D. Levine, "Online dominant and anomalous behavior detection in videos", pp. 2611-2618, 2013.

[7] M. Alvar, A. Torsello, A. Sanchez-Miralles, J.M. Armingol, "Abnormal behavior detection using dominant sets", Machine vision and applications, 25, pp. 1351-1368, 2014.

[8] T. Wang, J. Chen, H. Snoussi, "Online detection of abnormal events in video streams", Journal of Electrical and Computer Engineering, 2013, pp. 20-20, 2013.

[9] Y.S. Chong, Y.H. Tay, "Modeling representation of videos for anomaly detection using deep learning: A review", arXiv preprint arXiv:1505.00523, 2015.

[10] Yong Shean Chong and Yong Haur Tay. "Abnormal event detection in videos using spatiotemporal autoencoder". In International symposium on neural networks, pages 189–196. Springer, 2017.

[11] C. Chen, A. Heili, J.-M. Odobez, "A joint estimation of head and body orientation cues in surveillance video", IEEE, pp. 860-867, 2011.

[12] Z. Xu, I.W. Tsang, Y. Yang, Z. Ma, A.G. Hauptmann, "Event detection using multi-level relevance labels and multiple features", pp. 97-104, 2014.

[13] A. Nurhadiyatna, W. Jatmiko, B. Hardjono, A. Wibisono, I. Sina, P. Mursanto, "Background subtraction using gaussian mixture model enhanced by hole filling algorithm (gmmhf)", IEEE, pp. 4006-4011, 2014.

[14] S. Duffner, C. Garcia, "Pixeltrack: a fast adaptive algorithm for tracking non-rigid objects", pp. 2480-2487, 2013.

[15] J. Ning, L. Zhang, D. Zhang, W. Yu, "Joint registration and active contour segmentation for object tracking", IEEE transactions on circuits and systems for video technology, 23, pp. 1589-1597, 2013.

[16] Y. Zhang, H. Lu, L. Zhang, X. Ruan, "Combining motion and appearance cues for anomaly detection", Pattern Recognition, 51, pp. 443-452, 2016.

[17] F. Salo, M. Injadat, A.B. Nassif, A. Shami, A. Essex, "Data mining techniques in intrusion detection systems: A systematic literature review", IEEE Access, 6, pp. 56046-56058, 2018.

[18] H. Lu, H.S. Li, L. Chai, S.M. Fei, G.Y. Liu, "Multi-feature fusion based object detecting and tracking", Trans Tech Publ, pp. 1824-1828, 2012.

[19] F. Salo, M. Injadat, A. Moubayed, A.B. Nassif, A. Essex, "Clustering enabled classification using ensemble feature selection for intrusion detection", IEEE, pp. 276-281, 2019.

[20] M. Hasan, J. Choi, J. Neumann, A.K. Roy-Chowdhury, L.S. Davis, "Learning temporal regularity in video sequences", pp. 733-742, 2016.

[21] V. Reddy, C. Sanderson, B.C. Lovell, "Improved anomaly detection in crowded scenes via cell-based analysis of foreground speed, size and texture", IEEE, pp. 55-61, 2011.

[22] S. Sharma, S. Sebastian, "IoT based car accident detection and notification algorithm for general road accidents", International Journal of Electrical & Computer Engineering (2088-8708), 9, 2019.

[23] [Y. Kozlov, T. Weinkauf, Persistence1D: "Extracting and filtering minima and maxima of 1d functions", Accessed, 2015.

[24] Yiwei Lu, K Mahesh Kumar, Seyed shahabeddin Nabavi, and Yang Wang. "Future frame prediction using convolutional vrnn for anomaly detection". In 2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pages 1–8. IEEE, 2019.

[25] T. Xiao, C. Zhang, H. Zha, F. Wei, "Anomaly detection via local coordinate factorization and spatio-temporal pyramid", Springer, pp. 66-82, 2015.

[26] T. Wang, H. Snoussi, "Histograms of optical flow orientation for abnormal events detection", IEEE, pp. 45-52, 2013.

[27] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, W.-c. Woo, "Convolutional LSTM network: A machine learning approach for precipitation nowcasting", Advances in neural information processing systems, 28, 2015.

[28] V. Patraucean, A. Handa, R. Cipolla, "Spatio-temporal video autoencoder with differentiable memory", arXiv preprint arXiv:1511.06309, 2015.

[29] M. Sabokrou, M. Fathy, M. Hoseini, R. Klette, "Real-time anomaly detection and localization in crowded scenes", pp. 56-62, 2017.

[30] Adam. K, Abhishek Joshi and Vinay P Namboodiri. "Unsupervised synthesis of anomalies in videos: Transforming the normal". In 2019 International Joint Conference on Neural Networks (IJCNN), pages 1–8. IEEE, 2019.

[31] K.-W. Cheng, Y.-T. Chen, W.-H. Fang, "Gaussian process regression-based video anomaly detection and localization with hierarchical feature representation", IEEE Transactions on Image Processing, 24, pp. 5288-5301, 2015.

[32] Y. Cong, J. Yuan, J. Liu, "Sparse reconstruction cost for abnormal event detection", IEEE, pp. 3449-3456, 2011.

[33] Y. Hu, Y. Zhang, L. Davis, "Unsupervised abnormal crowd activity detection using semiparametric scan statistic", IEEE, pp. 767-774, 2013.

[34] V. Kaltsa, A. Briassouli, I. Kompatsiaris, L.J. Hadjileontiadis, M.G. Strintzis, "Swarm intelligence for detecting interesting events in crowded environments", IEEE transactions on image processing, 24, pp. 2153-2166, 2015.

[35] Kim. M, Nikos Komodakis and Spyros Gidaris. "Unsupervised representation learning by predicting image rotations". In International Conference on Learning Representations (ICLR), Vancouver, Canada, Apr. 2018.

[36] W. Li, V. Mahadevan, N. Vasconcelos, "Anomaly detection and localization in crowded scenes", IEEE transactions on pattern analysis and machine intelligence, 36, pp. 18-32, 2013.

[37] V. Saligrama, Z. Chen, "Video anomaly detection based on local statistical aggregates", IEEE, pp. 2112-2119, 2012.

[38] S. Wu, B.E. Moore, M. Shah, "Chaotic invariants of lagrangian particle trajectories for anomaly detection in crowded scenes", IEEE, pp. 2054-2060, 2010.

[39] C.F. Keras, GitHub, Seattle, WA, USA, 2015.

[40] J. Bergstra, O. Breuleux, F. Bastien, P. Lamblin, R. Pascanu, G. Desjardins, J. Turian, D. Warde-Farley, Y. Bengio, "Theano: a CPU and GPU math expression compiler", Austin, TX, pp. 1-7, 2014.

[41] T.T.D. Team, R. Al-Rfou, G. Alain, A. Almahairi, C. Angermueller, D. Bahdanau, N. Ballas, F. Bastien, J. Bayer, A. Belikov, "Theano: A Python framework for fast computation of mathematical expressions", arXiv preprint arXiv:1605.02688, 2016.

[42] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, "Tensorflow: a system for large-scale machine learning", Savannah, GA, USA, pp. 265-283, 2015.

[43] D. Xu, R. Song, X. Wu, N. Li, W. Feng, H. Qian, "Video anomaly detection based on a hierarchical activity discovery within spatio-temporal contexts", Neurocomputing, 143, pp. 144-152, 2014.