Anomaly Detection in Videos using Optical Flow and Convolutional Autoencoder

Article *in* IEEE Access · December 2019

DOI: 10.1109/ACCESS.2019.2960654

CITATIONS
3

READS
933

2 authors:

Elvan Duman
Mehmet Akif Ersoy University
12 PUBLICATIONS **11** CITATIONS

SEE PROFILE

Ayhan Erdem
Gazi University
65 PUBLICATIONS **98** CITATIONS

SEE PROFILE

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

SPECIAL SECTION ON INTEGRATIVE COMPUTER VISION AND MULTIMEDIA ANALYTICS

Received November 19, 2019, accepted December 12, 2019, date of publication December 18, 2019,
date of current version December 27, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2960654

Anomaly Detection in Videos Using Optical Flow and Convolutional Autoencoder

ELVAN DUMAN, (Student Member, IEEE), AND OSMAN AYHAN ERDEM
Department of Computer Engineering, Technology Faculty, Gazi University, 06500 Ankara, Turkey
Corresponding author: Elvan Duman (elvanduman@gazi.edu.tr)

ABSTRACT Today, public areas, such as airports, hospitals, city centers are monitored by surveillance systems. The widespread use of surveillance systems reduces security concerns while creating an amount of video data that cannot be examined by people in real-time. Therefore, the concept of automatic understanding of video activities has raised the standards of security camera systems. In this paper, we propose a framework (OF-ConvAE-LSTM) to detect anomalies using Convolutional Autoencoder and Convolutional Long Short-Term Memory in an unsupervised manner. Besides the deep learning model, the feature extraction stage based on dense optical flow is applied in the framework to obtain the velocity and direction information of foreground objects. The experiments were carried out on three popular public datasets consisting of Avenue, UCSD Ped1, and UCSD Peds2. The experimental results have shown that the proposed framework models the complex distribution of the pattern of regular motion changes with high accuracy. Besides, this method was observed to outperform state-of-the-art approaches based on unsupervised and semi-supervised deep learning models.

INDEX TERMS Abnormal event detection, convolutional autoencoder, long short-term memory, optical flow.

I. INTRODUCTION

In the last decade, the widespread use of surveillance systems has brought about massive amounts of video data. For this reason, there is a growing requirement for computer-aided tools performing the automatic analysis of these data. Especially automatic understanding of activities in videos with an unsupervised manner is a very critical problem for computer vision applications [1].

The classification of abnormal activities has attracted significant attention in the area of image processing [2]. In particular, automatic detection of abnormal activities, behaviors, or events in complex and crowded scenes is an important challenge because the definition of anomalous events in videos does not only depend on the context but also human-defined semantics. Furthermore, video data is hard to represent and model because of its high dimensionality, noise, and a wide variety of events. There are many fundamental limitations in anomaly detection of video surveillance systems. The main limitation is that there is a lack of a commonly accepted definition of anomaly because it varies significantly depending on the given scenario. For instance, running in a bank could be defined as an abnormal behavior while running at a traffic light could be normal. Another limitation in this area is to find out in which video frames anomalies occur and to localize regions that cause anomalies within these video clips. These limitations have made it difficult for computer systems to detect video anomalies.

The extraction of proper features plays an important role in anomaly detection methods to be capable of detecting a wide variety of anomalies [3]. The most popular feature extraction methods are based on spatial and temporal features such as Optical Flow-based [4], Histogram of Oriented Gradients-based [5], and trajectory-based [6]. In addition to feature extraction, learning models also have a remarkable effect on the efficiency of the approaches. Recently, Convolutional Neural Networks have been widely seen in image processing applications. There are some critical applications using CNNs such as object detection [7], face recognition [8], and edge detection [9], which is closely related to the application of the paper. In particular, CNNs have shown remarkable performance in anomaly detection for videos [10], [11]. The main advantage of CNN compared to its predecessors is that it automatically extracts the key features without any human effort. However, CNNs generally need strong-supervision, which is necessary for the training stage. Labeling data is highly

The associate editor coordinating the review of this manuscript and approving it for publication was Guitao Cao.

time-consuming and sometimes requires expert knowledge. Besides, it may also be difficult to find a set of data containing abnormalities in large sizes.

Recently, researchers aim to bridge the gap between the performance of supervised learning, which CNNs use in the training stage, and unsupervised learning. Learning valuable features by using large unlabeled datasets has been an area of intensive research with the advent of deep generative models which include convolutional autoencoder [12], variational autoencoder [13], generative adversarial network [14], convolutional long short-term memory [15], and their variations.

This paper presents a new framework in an unsupervised manner that benefited from dense optical flow-based handcrafted feature extraction, Convolutional autoencoder, and Convolutional LSTM. The pattern of apparent motion of foregrounds was obtained, such as velocities and movement directions for the input of the model. Then, the temporal and spatial patterns of normal activities were learned with the framework. Besides, while previous studies in the literature [16]–[18] have worked on grayscale video frames with a square aspect ratio such as 224x224, our framework used RGB video frames with an aspect ratio of 16:9, known as HD widescreen, as an input.

The rest of this paper is organized as follows: Section 2 provides the literature regarding the development of anomaly detection methods and state-of-the-art methods. Section 3 describes the details and key aspects of our framework and its stages. Section 4 discusses the experimental results. Finally, section 5 presents the conclusions of our study and suggestions for future work.

II. RELATED WORKS

Regarding the existing learning models, the approaches consider anomaly detection as a binary classification task, and activities are classified as either normal or abnormal. An extensive amount of literature has been devoted to the field of anomaly detection in videos, and researchers have published valuable survey articles on the literature [1], [3], [19], [20]. The vast majority of anomaly detection methods in videos engage a handcrafted feature extraction stage, followed by establishing a pattern model using videos containing only regular activities. Any activity that does not fit the learned model is considered as an anomaly. Previous studies in abnormal event detection in videos concentrated on handcrafted features and deep learning-based approaches.

Handcrafted approaches extract different types of motion information such as trajectories, optical flow, and histogram of gradient. Then they try to model normal or abnormal patterns of the scenes. Trajectory estimation is a widely used task in the feature extraction process. The trajectory-based models generally learn the normal pattern of trajectories and describe the dynamic information of moving objects. Visual tracking algorithms are employed to obtain the motion of moving objects in the videos. In a pioneering study [21], a statistical-based method, which learns from the long image sequences for event detection, was presented. For tracking multiple objects, Hu *et al.* [22] proposed a system that automatically learns motion patterns of moving objects and uses a fuzzy k-means based tracking algorithm. Jiang *et al.* [23] proposed a video event detection method based on Hidden Markov model (HMM) in an unsupervised learning manner. The model is capable of preventing overfitting via including a dynamic hierarchical process incorporated. In another study [24], unlike convolutional trajectory-based approaches, they combined the output of trajectory and pixel-based analysis. Therefore, the method is able to consider not only object trajectories but also the speed of the objects. As a result, trajectory-based approaches have good performance when the scenes are sparse because detecting and tracking moving objects are easy in sparse scenes. On the opposite, the success of the trajectory-based methods degrades in crowded scenes due to the difficulty of detecting and tracking objects.

Deep learning has been one of the most remarkable rising research areas due to bearing high potential to change the pattern of learning in almost every area of human life. Deep learning-based approaches have proved their success in many image processing and computer vision tasks as well as anomaly detection in videos. Researchers have recently proposed generative models of regular activity patterns with the consideration of the difficulty of finding abnormal activity patterns. Generative adversarial networks and especially convolutional neural networks are commonly used to build generative models. Convolutional neural network was initially used for anomaly detection to obtain both spatial and temporal features in the study [10]. However, CNNs were not built to learn temporal features in mind and are not a natural fit for videos. Convolutional Autoencoder is a good alternative to CNNs. Hasan *et al.* [12] proposed an approach that benefits from both fully connected autoencoder and fully convolutional feed-forward autoencoder to learn temporal regularity. The approach benefits from either state-of-the-art motion features or obtained features from autoencoder. To learn spatio-temporal representations better, Medel and Savakis [24] proposed a convolutional long-short-term memory network that is capable of encoding a video sequence. Another study [2] takes advantage of ConvAE and ConvLSTM together. Thus, the framework allows the extraction of spatial features and their temporal evolutions. In a similar study [25], The ConvAE + ConvLSTM model was trained by minimizing a Weighted Euclidean Loss. Principal Component Analysis (PCA) was used for segmentation of moving foregrounds from the video volume, and then the loss function was calculated with only extracted foregrounds. Thus, the influences of backgrounds were reduced.

Previous literature has shown that unsupervised deep learning models are prevalent and making progress in performance day by day. Especially, autoencoder based methods have come to the fore due to the easiness of its application and relatively reduced processing time for real-time surveillance systems.

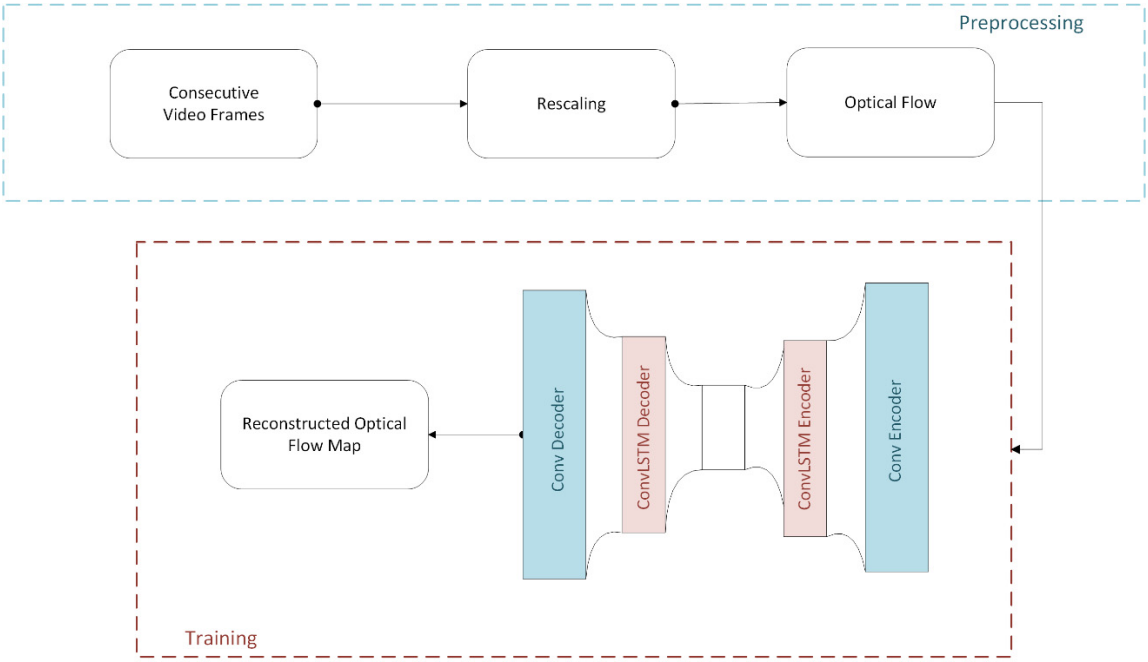

FIGURE 1. The architecture of the proposed framework: the preprocessing and training stage.

III. MATERIALS AND METHODS

In this study, a convolutional autoencoder method is employed to learn the pattern of normal activities in videos. The main idea of the framework is that the frames, which contain any abnormal event, give significantly different motion pattern than the normal frames. As an input to the encoder, dense optical flow maps are used. Then the network is trained with videos in which no abnormal event is included. After the training stage is properly done, the autoencoder can model the complex distribution of the pattern of normal motion changes. If an input video has an abnormal event, the model is expected to give a higher reconstruction error. Besides, the model was able to reconstruct optic flow maps for corresponding normal video volumes.

Our framework consists of three main stages. The first stage of the framework, called preprocessing, aims at extracting dense optical flow map of each frame. In the second stage, the convolutional autoencoder is used in order to obtain the spatial structure of each dense optical flow map volume. The last stage includes a convolutional long short-term memory network to learn the temporal patterns of encoded optical flow maps.

A. PREPROCESSING

The objective of the preprocessing step is to obtain acceptable input in terms of suitability to real-time application and the model. The preprocessing stage consists of two parts: rescaling and obtaining optical flow maps of each eight consecutive video frames. The universal video format of today is 16:9 [26], which is an aspect ratio with a width of 16 units and a height of 9 units. In our study, each frame was extracted from the dataset videos and rescaled to 320 x 180. In this way, we both reduced the dimension of the frames and kept the aspect ratio. At the last step of the preprocessing, the frames were converted to grayscale. Preprocessing flow is given in Fig. 1.

The inputs of the framework are formed by optical flow displacements areas between eight consecutive video frames. As the second step in the preprocessing, Farneback optical flow is employed to obtain dense optical flow maps. Therefore, the velocity and direction information of foregrounds objects are obtained. Farneback [27] proposed a dense optical flow algorithm based on modeling the neighbors of an observed pixel by quadratic polynomial basis. The term dense means optical flow is calculated for every single pixel in the frame. The cost resulting from the calculation of optical flow for each pixel can be easily solved with Farneback algorithm as it is linear. The principle is to represent the video signal in the neighborhood of each video frame by a 3D surface and identify the apparent motion of frame objects by finding the motion of pixels moving between two consecutive frames. For Farneback algorithm, quadratic polynomials expressed in a local coordinate system,

$$f(x) \sim x^T A x + b^T x + c \tag{1}$$

where A represents a symmetric matrix, b stands for a vector, and c is a scalar. The coefficients are computed from a weighted least squares fit to the signal values in the neighborhood. Polynomial for the first frame is given in (2), and

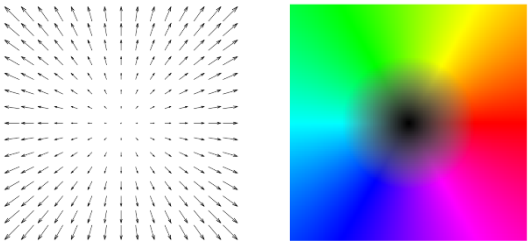

FIGURE 2. An illustration of the vector field and the optical flow coloring scheme.

the next frame with displacement is given in (3).

$$f_1(x) \sim x^T A_1 x + b_1^T x + c_1 \tag{2}$$

$$f_2(x) = f_1(x - d) \tag{3}$$

$$(x-d)^T A_1 (x-d) + b_1^T (x-d) + c_1 = x^T A_2 x + b_2^T x + c_2 \tag{4}$$

Displacement d is obtained by approximating the neighborhood polynomials on two consecutive video frames. The frame $f_1(x)$ is obtained at time t, and consecutive frame $f_2(x)$ is obtained at time $(t + dt)$.

Equating coefficients of (2) and (3) in the quadratic polynomials yields,

$$A_2 = A_1 \tag{5}$$

$$b_2 = b_1 - 2A_1 d \tag{6}$$

$$c_2 = d^T A_1 d - b_1^T x + c_1 \tag{7}$$

When at least A_1 is non-singular, (6) can be solved for the translation d.

$$2A_1 d = -(b_2 - b_1) \tag{8}$$

$$d = -\frac{1}{2} A_1^{-1} (b_2 - b_1) \tag{9}$$

An obtained displacement vector yields a motion vector corresponding to displacement. At the final stage of optical flow, displacement vectors are converted to color codes displayed in Fig. 2. The colors represent the direction of the optical flow vector, while the magnitude is converted into the color intensity. The dark colors indicate relatively low flow magnitudes, and green, blue, red, and other colors stand for a specific direction and flow magnitudes.

After the preprocessing step, optical flow maps of eight consecutive video frames are obtained. These optical flow maps would be the inputs of our deep learning model. The training stage will be explained in the next section.

B. DEEP LEARNING MODEL

Our deep learning model is shown in Fig. 1 as a training stage. The model consists of a spatial decoder and an encoder with convolutional LSTM layers. In this way, it is aimed to establish an unsupervised model in order to learn the motion pattern of normal behavior in the training phase. As known, the traditional autoencoder models are incapable of maintaining the spatial features of images [28]. Therefore, Masci *et al.* [29] proposed a convolutional autoencoder that learns the optimal filters for minimizing reconstruction error. Convolutional neural networks generally provide unsupervised learning and solve classification problems. In contrast, convolutional autoencoders provide reconstructed images by using back-propagation in an unsupervised manner and preserve the spatial relationship of video frames.

Spatial convolution maintains the spatial correlation between image areas by using square filters. These filters are used in convolution operations that perform dot products with local regions of a given image. Assume that we are given n x n square and m x m filter, the convolutional layer output will be (n-m+1) x (n-m+1). Besides filter properties, the parameters, such as number of layers, number of filters, size of filters, and stride values are needed to be specified for the learning performance of the model. The most important issue that restricts us is that increasing the number of filters requires extra computational time and memory. In this study, the proposed architecture was determined considering the needs and experimental studies. The structure of the proposed deep learning method is shown in Fig. 3.

For sequence modeling and learning, LSTM, as a special recurrent neural network architecture, has achieved remarkable results for temporal sequences in various studies [30]–[33]. The major advantage of the LSTM is that it has a memory cell in its structure [34].

The LSTM has achieved remarkable success in obtaining correlations between temporal events. ConvLSTM [34] allows LSTM to work on video data by substituting the matrix multiplication of LSTM with the convolutional operation. In this way, ConvLSTM captures both temporal and spatial information and provides superior performance with video frames comparing to fully connected LSTM [25]. As shown in Fig. 3, our framework consists of three ConvLSTM layers in order to learn temporal information of video frame sequences.

The ConvLSTM is expressed as follow [34]:

$$i_t = \sigma(w_{xi} * x_t + w_{hi} * h_{t-1} + b_i) \tag{10}$$

$$f_t = \sigma(w_{xf} * x_t + w_{hf} * h_{t-1} + b_f) \tag{11}$$

$$o_t = \sigma(w_{xo} * x_t + w_{ho} * h_{t-1} + b_o) \tag{12}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tanh(w_{xc} * x_t + w_{hc} * h_{t-1} + b_c) \tag{13}$$

$$h = o_t \odot \tanh(c_t) \tag{14}$$

where x_t and h_t refer to the network input and output at time t; i_t, f_t, o_t stand for the input gate, forget gate, and output gate, respectively. The temporal information is stored in the memory cell c_t. The weights of the convolutional filters are represented with w_{xi}, w_{hi}, w_{xf}, w_{hf}, w_{xo}, w_{ho}, w_{xc}, and w_{hc}. σ and b represent the sigmoid activation function and bias, respectively. The operation $\odot$ stands for multiplication, and $*$ stands for convolutional operation.

C. REGULARITY SCORE

For performance assessment, the optical flow map volumes are compared, which are the inputs of our model obtained

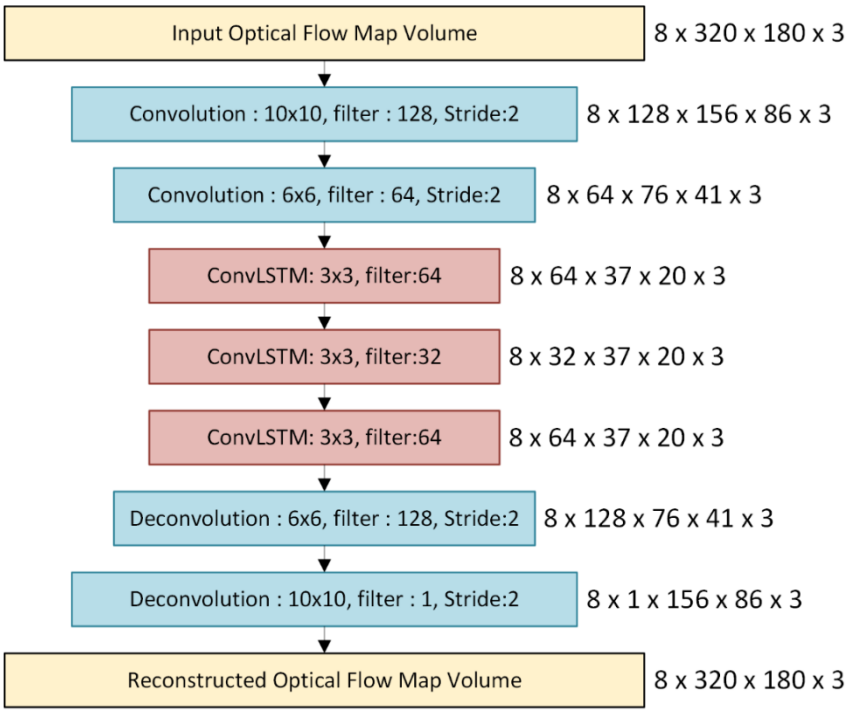

FIGURE 3. The detailed convolutional network architecture of the proposed framework. It has a spatial decoder and encoder with convolutional LSTM layers. The input is an optical flow map of eight sequence frames. The dimensions given at the right side represent the output size of the corresponding layer. There are three ConvLSTM layers to learn temporal motion correlations.

after the preprocessing stage, with reconstructed optical flow map volumes, which are the outputs of our model. After the training stage, a reconstruction error between an optical flow volume and corresponding reconstructed optical flow volume for each eight consecutive video frames is computed according to the multi-scale structural similarity (MS-SSIM) [35] quality metric. State-of-the-art anomaly detection methods in videos generally use pixel-wise the mean squared error (MSE), which is derived by squared intensity differences of an input image and reconstructed image. MS-SSIM provides the similarity map in the pixel domain. For any two optical flow map input x and output y, MS-SSIM models the similarity between them as three complementary components, which are structural similarity, contrast similarity, and luminance similarity. The mathematical notations of the metric were given in the study [35]. We compute the anomaly score of an optical flow volume (x) and corresponding reconstructed optical flow volume $f_w(x)$ as follows:

$$a\,(x) = 1 - (MS - SSIM(x, f_w(x)) \tag{15}$$

f_w denotes the weight obtained from the model. Anomaly score of volume $a(x)$ is obtained by summing up all the pixel-wise errors. Finally, the regularity score of $r(x)$ is derived by scaling to a range of the anomaly score to 0 to 1, and subtracting normalized anomaly score from 1:

$$r\,(x) = 1 - \frac{a\,(x) - min_t a(t)}{max_t a\,(t) - min_t a(t)} \tag{16}$$

We also used the Persistence1D [36] algorithm to avoid the local minima problem in the regularity score. We assume that an event must take at least 48 frames to occur to deal with the local minima problem. Thus, the success of the method is increased in the frame-level anomaly detection.

IV. EXPERIMENTS AND ANALYSIS

A. DATASETS

In order to conduct a comprehensive evaluation, we worked on three well-known datasets: Avenue [37], The UCSD Anomaly Detection Dataset [38]: Ped1, and Ped2. The datasets included both train and test videos. Training processes with the train videos containing only regular events were set up for each dataset. The models were evaluated with the test videos containing both normal and abnormal events.

Avenue dataset consists of 16 training and 21 test videos acquired with a stationary camera. The total frame number is 30652, and the duration of videos is equal or less than one minute. The average video length of videos was 37.75 seconds for training, and 28.6 seconds for testing. Training videos contain pedestrians at the entrance of the subway. The test dataset also includes regular events like a pedestrian walking to the subway entrance, as well as abnormal actions such as walking and running in the opposite direction. The challenges in Avenue dataset are some crowded scenes, camera shakes in one of the test videos, and some normal patterns that rarely come out in training data.

The UCSD Dataset was obtained with a stationary camera mounted at a high distance, overlooking

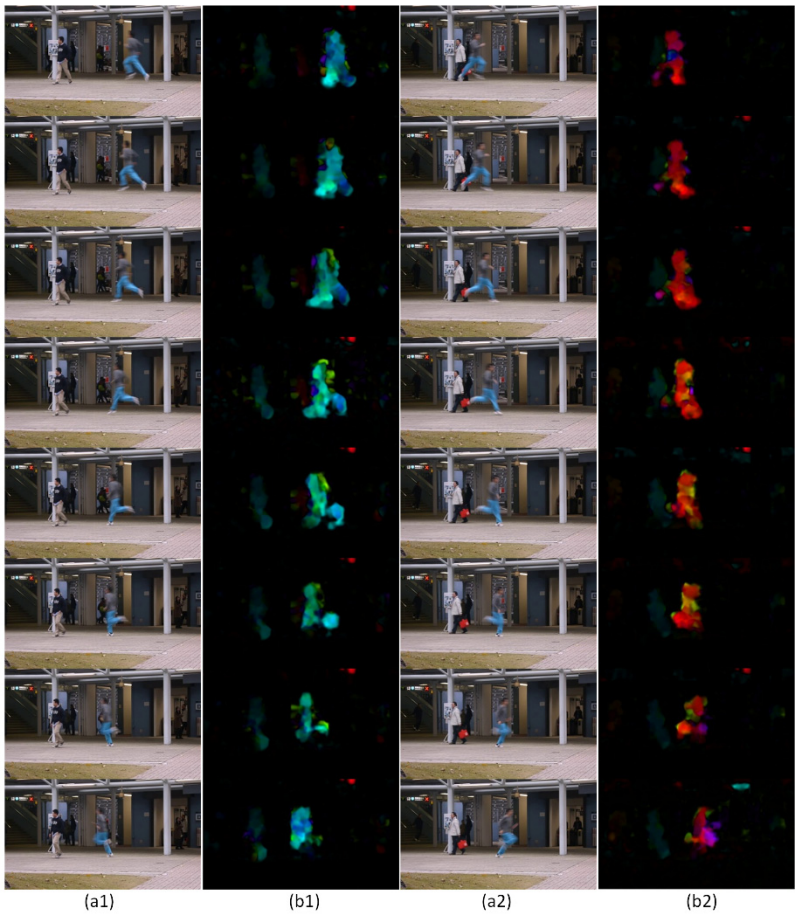

FIGURE 4. An example illustration of model input: eight consecutive frames and its optical flow map.

pedestrian walkways. The density of the scenes varies from sparse to quite a crowd. While the training videos contain only pedestrians, the test videos contain some abnormalities which occur due to either the circulation of non-pedestrian objects in the scenes or anomalous people's motion patterns. All videos are equal in length and consist of 200 frames. The Ped1 dataset consists of 32 training and 36 test video. The Ped2 consists of 16 training and 12 test videos.

B. IMPLEMENTATION DETAILS

We used Adam optimizer to optimize the model with the hyperbolic tangent as the activation function. The learning rate was automatically adjusted in accordance with the update history of model weights. The batch size was set as 64, and each training volume was trained up to 150 epochs. However, when the reconstruction lost does not decrease for 20 epochs, the training stage is automatically terminated. For optical flow, pyr_scale, window size, iterations, poly_n, and poly_sigma were selected as 0.5, 15, 3, 5, and 1.1, respectively.

Training and evaluation of the model were performed on a personal computer running 64-bit Windows 10 Pro operating system with an Intel Core i7-7700K CPU and 16 gigabytes DDR4 RAM. Graphics card was GeForce GTX 1070 with 256 bit 1920 CUDA cores and 8 GB GDRR5 RAM. Implementations of the models were written in python 3.6 version with using CUDA 9.

C. PERFORMANCE METRICS

In evaluating the performance of anomaly detection methods in videos, Receiver Operating Characteristic (ROC) curve and area under the curve (AUC) metrics were used. Using the AUC metric, the detection performance of our framework was quantitatively evaluated. The definitions of performance measure for comparisons are as follows:

- True positive (TP): number of abnormal activities correctly predicted
- True negative (TN): number of normal activities correctly predicted
- False positive (FP): number of normal activities incorrectly predicted as positive
- False negative (FN): number of positive samples incorrectly predicted as negative

Precision is a measure of the purity of all the samples labeled as positive and calculated by TP/(TP+FP). Recall is

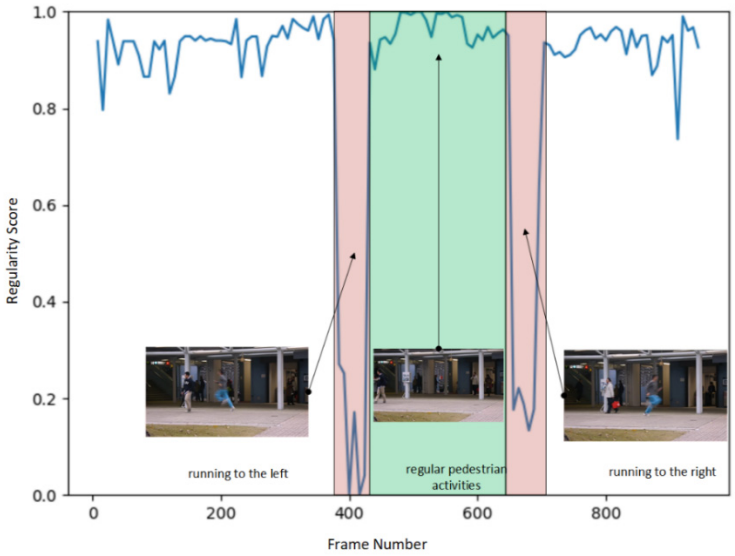

FIGURE 5. Experimental regularity score of video clip 4 from the test videos of avenue dataset. X-axes and y-axes denote regularity score and frame number of video clip 4, respectively. The fact that the regularity score approaches zero indicates the presence of an anomaly in the relevant frames.

a measure of the completeness of samples labeled as positive and calculated by TP/(TP+FN). ROC is a plot of the trade-off False Positive Rate (FPR) to True Positive Rate (TPR). FPR is calculated by FP/(FP+TN), and TPR is calculated by TP/(TP+FN). ROC curve provides a way to evaluate the model at all thresholds of score returned by a classifier. AUC is a common metric for assessing ROC curves, and it is the area under a ROC curve. The AUC degrades the curve to a single number in order to make the result easier to compare a considerable number of methods. The AUC value ranges from 0 to 1, and an AUC of 1 represents the best classification result. Further information and detailed definitions of the evaluation metrics can be found in [39].

D. EXPERIMENTAL RESULTS

The illustration of an example of model input is given in the second and fourth columns in Fig. 4. Each frame was extracted from the dataset videos and rescaled to 320 x 180. Then, dense optical flow of each frame was computed. The input of the model was defined as eight consecutive optical flow maps in Fig. 4 (b1) and Fig. 4 (b2).

The regularity score provides information about the possibility of abnormal activity in the video frames. Frames containing abnormal activities are expected to give a low regularity score, whereas frames containing normal activities are expected to give a high regularity score. Fig. 5. illustrates the regularity scores on a sample from the Avenue dataset. In the figure, red areas represent the anomalies obtained from the ground truth. The yellow area is an example of normal activity in video clip 4. The anomalies in video clip 4, running to the left and running to the right, can be observed as they produce low regularity scores with strong downward spikes. This situation indicates that the evaluated video frames contain abnormal conditions. As shown in Fig. 5, regular pedestrian activities give high regularity scores, and these scores approach 1.

We provide frame-level and event-level performance comparison for the datasets. The effectiveness of the proposed framework was tested by comparing our model with six different approaches based on autoencoder. These are ConvAE [12], ST-AE [2], ConvLSTM-AE [40], Two-Stream R-ConvVAE [41] and WCAE-LSTM [25], and STAN [42]. ConvAE [12] benefits from both fully connected autoencoder with trajectory-based handcrafted spatio-temporal features and convolutional autoencoder. The main limitation of the ConvAE is directly affected by the success of the trajectory-based feature extraction stage. It is because the success of the trajectory-based methods decreases in crowded scenes due to the difficulty of detecting and tracking objects. ST-AE [2] proposed spatio-temporal architecture, which leverages Convolutional AE and Convolutional LSTM without using any handcrafted local features. In this way, the method captures the spatial and the temporal features of videos with fast training and testing time. However, the fact that ST-AE does not use any handcrafted features as an input has reduced the performance of the method. ConvLSTM-AE [40] integrates Convolutional Neural Network and ConvLSTM with autoencoder in order to encode the pattern of normal activities. Two-Stream R-ConvVAE [41] aimed to model regular scenes by using two-stream recurrent variational autoencoder in a semi-supervised learning manner. WCAE-LSTM [25] proposed a weighted ConvAE and LSTM in order to encode spatial and temporal information of video frames. WCAE-LSTM [25] proposed a weighted Euclidean loss in order to extract foreground objects. In this way, the method was aimed to concentrate on moving objects without being affected by

TABLE 1. Frame-level performance comparison of state-of-the-art approaches and our framework. UAC illustrates the detection ability of the methods. AUC ranges in value from 0 to 1, and higher AUC is desired. A model of which the predictions are 100% true has an AUC of 1.

Method	Avenue	Ped1	Ped2
ConvAE[12]	75.2%	81.0%	90.0%
ST-AE [2]	80.3%	89.9%	87.4%
ConvLSTM-AE [40]	77.0%	75.5%	88.1%
Two-Stream R-ConvVAE [41]	79.6%	75.0%	91.7%
WCAE-LSTM[25]	85.7%	85.1%	92.6%
STAN[42]	87.2%	82.1%	**96.5%**
Our Framework	**89.5%**	**92.4%**	92.9%

TABLE 2. Information about the datasets and frame-level experimental results of our framework: TP: True positive; FP: False positive; TN: True negative; FN: False negative.

Database	Total Frame	Regular Frame	Anomalous Frame	TP	FP	TN	FN
Avenue	15324	11504	3820	3750	72	11417	87
UCSD Ped1	7200	3195	4005	3931	74	3143	38
UCSD Ped2	2010	374	1636	1568	68	358	16

TABLE 3. Event-level performance comparison of state-of-the-art approaches and our framework.

Database	Anomalous Event	Correct Detection/ False Alarm				
		ST-AE [2]	ConvAE[12]	TS-R-ConvVAE[41]	STAN[42]	Our Framework
Avenue	47	43/8	45/4	34/6	N/A	45/2
UCSD Ped1	40	36/11	38/6	38/6	37/3	38/2
UCSD Ped2	12	12/3	12/1	12/0	12/0	12/2

TABLE 4. Run-time details of preprocessing and testing stages.

Processing Unit	Preprocessing Stage	Testing Stage	Total Time
CPU	0.016 sec.	0.235 sec.	0,251 sec.(~4fps)
GPU	0.016 sec.	0.008 sec.	0,024 sec (~42.5 fps)

backgrounds. However, segmented foreground objects only provide information about the presence and shape of moving objects. Whereas, optical flow provides rate and direction of motions, alongside presence and shape information of moving objects. STAN [42] proposed a generative spatio-temporal adversarial network model to reveal whether the video frame sequence is real or fake. The model consists of a spatio-temporal generator and a spatio-temporal discriminator in order to efficiently represent spatio-temporal features of normal activity patterns. The generator generates inter-frames, and through the adversarial learning, the discriminator is trained to distinguish real frames from the frames generated by the generator. The advantage of STAN [42] is that it gives more robust results in complex scenes with frequent occlusions than our method. Despite the capabilities of the proposed methods, our framework is superior because it is less affected by the backgrounds and utilizes the velocity and direction information of foreground objects sufficiently.

Table 1 illustrates the frame-level performance comparison of area under ROC curve (AUC) between state-of-the-art methods and our framework for UCSD and Avenue datasets. According to the results, our framework is comparable to the state-of-the-art based on autoencoder. The proposed framework provided the best performance in anomaly detection for the Avenue and the Ped1 datasets with the 89.5% and 92.4% AUC scores, respectively. However, the AUC result of the STAN [42] is better than the proposed method for the Ped2 dataset. One of the reasons is the difficulty of our method in modeling activities that take place very far from the camera. Another reason may be that it is very difficult to get accurate and stable results due to the size of the Ped2 dataset, which contains only 16 training and 12 test videos.

The detailed information about the datasets and anomaly detection performance of our framework such as the total frames of the datasets and their frame numbers with anomalous activities, frame numbers with normal activities, and the number of true positives, false positives, true negatives,

and false negatives are provided in Table 2. Event-level performance comparison of the methods is also presented in Table 3. We got better results than state-of-the-art methods for the Avenue and Ped1 datasets. For the Avenue dataset, our framework could not detect only two running events. It could be because the running events were among the pedestrians and quite far away from the camera. We detected an equal number of anomalous events with the studies [12], [41]; however, the framework gave less false alarms compared to the studies. Despite detecting all anomalous of the Ped2 dataset, the false alarm rate of our framework is higher than the studies [12], [41], [42]. As a result, our framework has shown superior performance to the state-of-the-arts, considering the average accuracy for all datasets.

Run-time analysis on our framework for both CPU (Intel Core i7-7700k 4.20GHz) and GPU (NVIDIA GeForce GTX 10170) is given in Table 4. For CPU, the framework can handle approximately 4 frames; besides, the GPU process approximately 42.5 frames per second. This means that the graphics card runs about 10 times faster than the CPU.

V. CONCLUSION

A new framework was introduced to detect anomalies in videos based on an unsupervised generative model. The framework generates reconstructed dense optical flow maps and reconstruction error between input optical flow maps and corresponding reconstructed dense optical flow maps. In this paper, the success of the method in modeling normal behavior patterns and detecting abnormal behaviors was addressed. We experimentally evaluated our model on the datasets, and the results indicated that the framework is effective in anomaly detection in videos. For future study, we are planning to investigate how to obtain interaction forces between foreground objects in addition to optical flow information without reducing model interest for real-time applications.

ACKNOWLEDGMENT

The authors thank Gazi University Academic Writing Center for providing English-language assistance and proofreading this article.

REFERENCES

[1] 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, p. 36, 2018.
[2] Y. S. Chong and Y. H. Tay, "Abnormal event detection in videos using spatiotemporal autoencoder," in *Proc. Int. Symp. Neural Netw.* Cham, Switzerland: Springer, 2017, pp. 189–196.
[3] C. Dhiman and D. K. Vishwakarma, "A review of state-of-the-art techniques for abnormal human activity recognition," *Eng. Appl. Artif. Intell.*, vol. 77, pp. 21–45, Jan. 2018.
[4] T. Qasim and N. Bhatti, "A low dimensional descriptor for detection of anomalies in crowd videos," *Math. Comput. Simul.*, vol. 166, pp. 245–252, Dec. 2019.
[5] X. Hu, Y. Huang, Q. Duan, W. Ci, J. Dai, and H. Yang, "Abnormal event detection in crowded scenes using histogram of oriented contextual gradient descriptor," *EURASIP J. Adv. Signal Process.*, vol. 2018, no. 1, p. 54, 2018.
[6] Y. Shi, Y. Tian, Y. Wang, and T. Huang, "Sequential deep trajectory descriptor for action recognition with three-stream CNN," *IEEE Trans. Multimedia*, vol. 19, no. 7, pp. 1510–1520, Jul. 2017.
[7] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2012, pp. 1097–1105.
[8] H. Li, Z. Lin, X. Shen, J. Brandt, and G. Hua, "A convolutional neural network cascade for face detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 5325–5334.
[9] S. Xie and Z. Tu, "Holistically-nested edge detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1395–1403.
[10] S. Zhou, W. Shen, D. Zeng, M. Fang, Y. Wei, and Z. Zhang, "Spatial–temporal convolutional neural networks for anomaly detection and localization in crowded scenes," *Signal Process., Image Commun.*, vol. 47, pp. 358–368, Sep. 2016.
[11] L. Ding, W. Fang, H. Luo, L. Peter, B. Zhong, and X. Ouyang, "A deep hybrid learning model to detect unsafe behavior: Integrating convolution neural networks and long short-term memory," *Autom. Construct.*, vol. 86, pp. 118–124, Feb. 2018.
[12] 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.*, Jun. 2016, pp. 733–742.
[13] D. P. Kingma and M. Welling, "Stochastic gradient VB and the variational auto-encoder," in *Proc. 2nd Int. Conf. Learn. Represent. (ICLR)*, 2014.
[14] I. Goodfellow, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.
[15] T. N. Sainath, O. Vinyals, A. Senior, and H. Sak, "Convolutional, long short-term memory, fully connected deep neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2015, pp. 4580–4584.
[16] J. Fu, W. Fan, and N. Bouguila, "A novel approach for anomaly event detection in videos based on autoencoders and SE networks," in *Proc. 9th Int. Symp. Signal, Image, Video Commun. (ISIVC)*, Nov. 2018, pp. 179–184.
[17] Y. Zhao, B. Deng, C. Shen, Y. Liu, H. Lu, and X.-S. Hua, "Spatio-temporal autoencoder for video anomaly detection," in *Proc. 25th ACM Int. Conf. Multimedia*, 2017, pp. 1933–1941.
[18] J. R. Medel and A. Savakis, "Anomaly detection in video using predictive convolutional long short-term memory networks," Dec. 2016, *arXiv:1612.00390*. [Online]. Available: https://arxiv.org/abs/1612.00390
[19] A. B. Mabrouk and E. Zagrouba, "Abnormal behavior recognition for intelligent video surveillance systems: A review," *Expert Syst. Appl.*, vol. 91, pp. 480–491, Jan. 2018.
[20] T. Wang, "Abnormal event detection based on analysis of movement information of video sequence," *Optik*, vol. 152, pp. 50–60, Jan. 2018.
[21] N. Johnson and D. Hogg, "Learning the distribution of object trajectories for event recognition," *Image Vis. Comput.*, vol. 14, no. 8, pp. 609–615, 1996.
[22] W. Hu, X. Xiao, Z. Fu, D. Xie, T. Tan, and S. Maybank, "A system for learning statistical motion patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 9, pp. 1450–1464, Sep. 2006.
[23] F. Jiang, Y. Wu, and A. K. Katsaggelos, "A dynamic hierarchical clustering method for trajectory-based unusual video event detection," *IEEE Trans. Image Process.*, vol. 18, no. 4, pp. 907–913, Apr. 2009.
[24] S. Coşar, G. Donatiello, V. Bogorny, C. Garate, L. O. Alvares, and F. Brémond, "Toward abnormal trajectory and event detection in video surveillance," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 3, pp. 683–695, Mar. 2017.
[25] B. Yang, J. Cao, R. Ni, and L. Zou, "Anomaly detection in moving crowds through spatiotemporal autoencoding and additional attention," *Adv. Multimedia*, vol. 2018, pp. 1–8, Sep. 2018, Art. no. 2087574, doi: 10.1155/2018/2087574.
[26] H. Zeng, L. Li, Z. Cao, and L. Zhang, "Reliable and efficient image cropping: A grid anchor based approach," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 5949–5957.
[27] G. Farnebäck, "Two-frame motion estimation based on polynomial expansion," in *Proc. Scand. Conf. Image Anal.* Berlin, Germany: Springer, 2003, pp. 363–370.
[28] S. Mei, Y. Wang, and G. Wen, "Automatic fabric defect detection with a multi-scale convolutional denoising autoencoder network model," *Sensors*, vol. 18, no. 4, p. 1064, Apr. 2018.
[29] J. Masci, U. Meier, D. Cireşan, and J. Schmidhuber, "Stacked convolutional auto-encoders for hierarchical feature extraction," in *Proc. Int. Conf. Artif. Neural Netw.* Berlin, Germany: Springer, 2011, pp. 52–59.
[30] A. Graves, "Generating sequences with recurrent neural networks," Aug. 2013, *arXiv:1308.0850*. [Online]. Available: https://arxiv.org/abs/1308.0850

[31] A. Graves, A.-R. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2013, pp. 6645–6649.
[32] F. J. Ordóñez and D. Roggen, "Deep convolutional and LSTM recurrent neural networks for multimodal wearable activity recognition," *Sensors*, vol. 16, no. 1, p. 115, 2016.
[33] R. S. Arslan and N. Barışçı, "Development of output correction methodology for long short term memory-based speech recognition," *Sustainability*, vol. 11, no. 15, p. 4250, 2019.
[34] S. Xingjian, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-C. Woo, "Convolutional LSTM network: A machine learning approach for precipitation nowcasting," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 802–810.
[35] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *Proc. 37th Asilomar Conf. Signals, Syst. Comput.*, vol. 2, Nov. 2003, pp. 1398–1402.
[36] Y. Kozlov and T. Weinkauf. *Persistence1D: Extracting and Filtering Minima and Maxima of 1D Functions*. Accessed: 2015. pp. 1–11. [Online]. Available: http://people.mpi-inf.mpg.de/weinkauf/notes/persistence1d.html
[37] C. Lu, J. Shi, and J. Jia, "Abnormal event detection at 150 FPS in MATLAB," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 2720–2727.
[38] V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos, "Anomaly detection in crowded scenes," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 1975–1981.
[39] D. M. Powers, "Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation," *J. Mach. Learn. Technol.*, vol. 2, no. 1, pp. 37–63, 2011.
[40] W. Luo, W. Liu, and S. Gao, "Remembering history with convolutional LSTM for anomaly detection," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2017, pp. 439–444.
[41] S. Yan, J. S. Smith, W. Lu, and B. Zhang, "Abnormal event detection from videos using a two-stream recurrent variational autoencoder," *IEEE Trans. Cogn. Developmental Syst.*, to be published.
[42] S. Lee, H. G. Kim, and Y. M. Ro, "STAN: Spatio-temporal adversarial networks for abnormal event detection," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 1323–1327.

ELVAN DUMAN received the B.S. degree in electrical and computer education from Mugla University, the B.S. degree in computer engineering from Gazi University, Ankara, Turkey, and the M.S. degree in computer engineering from Kırıkkale University, Turkey, in 2012. He is currently pursuing the Ph.D. degree in computer engineering with Gazi University. He has been a Research Assistant with the Department of Computer Engineering, Gazi University, since 2012. His research interests include image processing, computer vision, and deep learning.

OSMAN AYHAN ERDEM received the B.S., M.S., and Ph.D. degrees from the Institute of Science and Technology, Gazi University, Turkey. In 1990, he has attended the English Language Education Program at Indiana University, USA. He finished the Technology of Computing Education at Purdue University. He is currently a Professor with the Department of Computer Engineering, Gazi University. He has published books and numerous articles in national and international journals. His current research interests inlcude programming languages, computer vision, computer systems, and computer networks.