Human-Machine Cooperative Video Anomaly Detection

FAN YANG and ZHIWEN YU, Northwestern Polytechnical University, China
LIMING CHEN, Ulster University, United Kingdom
JIAXI GU, QINGYANG LI, and BIN GUO, Northwestern Polytechnical University, China

It is still a challenge to detect anomalous events in video sequences in the field of computer vision due to heavy object occlusions, varying crowded densities and complex situations. To address this, we propose a novel human-machine cooperative approach which uses human feedback on anomaly confirmation to inform and enhance video anomaly detection. Specifically, we analyze the spatio-temporal characteristics of sequential frames of a video from the appearance and motion perspective from which spatial and temporal features are identified and extracted. We then develop a convolutional autoencoder neural network to compute an abnormal score based on reconstruction errors. In this process, a group of experts will provide human feedback to a certain proportion of classified frames to be incorporated into the model, and also the final judgment for the event anomalies for training and classification. The proposed approach is evaluated on 3 publicly available surveillance datasets, showing improved accuracy and competitive performance (93.7% AUC) with respect to the best performance (90.6% AUC) of the state-of-the-art approaches. The approach has not been previously seen to the best of our knowledge.

CCS Concepts: • **Human-centered computing** → *Collaborative interaction*; • **Computing methodologies** → *Object detection.*

Additional Key Words and Phrases: human-machine; anomaly detection; autoencoder; video frame

ACM Reference Format:
Fan Yang, Zhiwen Yu, Liming Chen, Jiaxi Gu, Qingyang Li, and Bin Guo. 2020. Human-Machine Cooperative Video Anomaly Detection. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW3, Article 274 (December 2020), 18 pages. https://doi.org/10.1145/3434183

1 INTRODUCTION

In the past few years, the prevalence of video surveillance systems in public areas has led to a growing demand on advanced methods for efficient analysis of video streams. The deployed camera systems not only record what happened over time but also offer an invisible power to assure people security to some extent. From a large amount of surveillance videos, how to automatically and efficiently find possible anomalies or objects of interest has become a challenging task. Anomaly detection is becoming a significant and critical branch in computer vision research community. Different from supervised video analysis problem such as action recognition [18] and events detection [25], video anomaly detection is not applicable by training a model on positive (anomalous) samples, due to the sparsity of anomalous events and a large variety of different anomalous events. Thus, a reasonable method relies on data which only contains normal videos. Learning

Authors' addresses: Fan Yang, yang-fan@mail.nwpu.edu.cn; Zhiwen Yu, zhiwenyu@nwpu.edu.cn, Northwestern Polytechnical University, 1 Dongxiang Road, Chang'an District, Xi'an, China, 710129; Liming Chen, Ulster University, Shore Road, Newtownabbey, Belfast, United Kingdom, l.chen@ulster.ac.uk; Jiaxi Gu; Qingyang Li; Bin Guo, Northwestern Polytechnical University, Xi'an, China.

2573-0142/2020/12-ART274 $15.00
https://doi.org/10.1145/3434183

the feature representation of regular activities is an unsupervised learning problem [27]. Some previous anomaly detection works [1, 12, 23] focus on modeling spatio-temporal event patterns corresponding to appearances and motions of local 2D image patches or 3D video cubes using hand-crafted features, such as, histogram of oriented gradients (HOG) [13], histogram of optical flow (HOF) [11], 3D spatio-temporal gradient [4]. However, due to the limited representation capability of hand-crafted features, this category of approaches are not applicable to complex video surveillance scenes.

Recently, deep learning approaches have shown significant advantages for various computer vision tasks, such as object classification [19], object detection [26]. While most studies focus on supervised learning tasks and Convolutional Neural Networks, unsupervised approaches have also gained popularity because of the fact that they can extract rich features as well as some hidden nature via multi-layer nonlinear transformations. In particular, autoencoder networks [29] have been investigated to address video anomaly detection problems [10, 16, 33]. However, these methods extract features from fully-connected autoencoder or convolutional autoencoder, without leveraging temporal dimensions, which is essential for recognizing video outliers. We use a spatio-temporal model with video frame sequences and corresponding histogram of optical flow as input to obtain both spatial and temporal aspects of features. No matter how robust the machine based approach is, there are still a certain number of false alarms and missed anomalies. In particular, some crowded scenes where people and objects occlude each other bring more errors in anomaly detection.

In order to improve the performance of video anomaly detection, one may need to collect more samples containing anomalies for constructing a robust model, but there are still some anomaly scenarios such as shielded targets which are easy for humans but very hard for a machine to detect. Inspired by this observation, we add human (domain experts) feedback including labeling, assessing and correcting in our anomaly detection framework. In this way, we can use the label information for the processing model to classify those anomalous events which are hard to be identified. Based on the above ideas, in this paper we propose a novel approach for anomaly detection in complex video surveillance scenes by learning discriminative features in a human-in-the-loop supervised manner adopting convolutional autoencoders (CAE). Fig. 1 shows an overview of the proposed method, called Human-Machine Cooperation Framework (HMCF). Our approach is based on a novel fusion scheme (integrating both traditional early fusion and late fusion strategies) for combining low-level features of appearance and motion. Specifically, in the first phase, individual video frames and their corresponding optical flow fields are provided as input to a common deep autoencoder network, to learn appearance and motion features. In the second phase, experts respond to requests from the model to confirm whether or not a video frame contains anomalous objects. Anomaly detection is conducted by computing reconstruction error with lower errors indicating regular frames while higher errors irregular frames. The proposed HMCF is evaluated using three video surveillance datasets and the evaluation results are compared with the performance of several state-of-the-art methods. Our experimental results clearly demonstrate the effectiveness of proposed approach. In general, we make the following contributions:

(1) Conceive and create the Human-Machine Cooperation Framework (HMCF) for video anomaly detection and analysis which is the first attempt of using human-machine cooperation to improve the efficiency of video anomaly detection to the best of our knowledge.
(2) Develop a convolutional autoencoder model and corresponding methods to extract spatio-temporal features of the video sequences and compute the reconstruction errors between input and output frames, which provides fine-grained frame-level detection features.

(3) Develop a new mechanism and associated intuitive user interface to facilitate interactive human-machine cooperation on hybrid anomaly detection. Experiments show that the cooperation mechanism is effective for difficult-to-detect events and avoids the need of retraining detection model.

2 RELATED WORK

Our work is mainly relevant to the following two areas of research: video anomaly detection, and interactive machine learning.

Video anomaly detection. Generally, an anomaly is defined in a specific situation, and opposite to a large number of normal objects, it might be a person riding a bicycle, a person walking around. There are different anomalies in different datasets, so it is impossible to train all kinds of anomalous samples. The rationale of some existing approaches is to learn an object model from normal videos, and then detect abnormal events as samples which disagree with the normal event model. According to the processing pattern, existing algorithms for anomaly detection can be divided into two categories: 1) Hand-crafted feature techniques. Features extracted from some descriptors what human expected or preconceived. 2) Deep learning based methods. Features are automatically extracted by neural network. For the first category, researchers have investigated on the trajectories which are extracted in advance for moving objects. By exploring potential rules among normal trajectories, abnormal events are identified as ones which disobey these rules [8]. An object can be viewed as an anomaly if it does not follow learned normal trajectories, however, it does not work in some scenarios, such as disability to efficiently handle occlusions, and having high complexity in crowded scenes. For example, authors in [27] extracted multiple features based on trajectories, speed and acceleration. In their methods, each feature is applied with a clustering algorithm, and the final clustering result is obtained by taking clusters from all features into account. The clusters with few members and samples far away from these cluster centers are treated as anomalies. In order to deal with the occlusion and segmentation problems, Wu et al. [31] proposed a Lagrangian particle dynamics approach, and extracted chaotic invariant features from representative trajectories.

Although the trajectory-based detection methods are improved to some extent, their performance is still not satisfactory. In order to avoid the weaknesses, some low-level features (spatial and temporal) are extracted by some classical descriptors, such as, histogram of oriented gradients (HOG) or histogram of optical flow (HOF) [21, 23]. In [23], Mahadevan et al. proposed a joint detector for detecting temporal and spatial anomalies, they made use of a mixture of dynamic textures (MDT) for representing the video and fitting a Gaussian mixture model to features. Cong et al. [12] and Lu et al. [22] learned an over-complete normal basis set from training data, and they introduced a cost for sparse reconstruction of a testing patch for detecting anomaly patches.

For the second category, deep learning has achieved substantial ascension in many computer vision tasks. For instance, event features are extracted from 2D image patches or 3D video blocks, it is proposed to use spatio-temporal features such as optical flow or gradients. Adam et al. [1] used an exponential distribution for modeling the histograms of optical flow in local regions. Xu et al. [33] proposed a method for detecting anomalies based on denoise convolutional autoencoder and a fusion scheme of early and later pattern to attain a better effect. Sabokrou et al. [27] designed models for normal events based on a set of representative features which were learned by autoencoders, which was proved to be effective to recognize unusual events. In [16] the authors investigated the regular temporal features from the trajectory-based method by HOF and HOG, and another way was extracting high level features by a deep autoencoder neural network. Chong et al. [10] proposed a CNN combined with LSTM based on autoencoder to obtain spatio-temporal features and compute its reconstruction error. In [34] Zhao designed a 3D spatio-temporal autoencoder to detect the nearly future anomaly. As a matter of fact, humans are very competent to intuitively combine different features, such as motion and appearance features, in order to understand the meaning of anomaly video sequences.

Interactive Machine Learning (IML). IML incorporates human feedback in the model training process to create better ML models [14], and has become a hot area of research [3, 15]. For example, using user-provided examples, Amershi et al. [3] enabled an algorithm to learn about new friend groups on social media. In [15], Fogarty et al. allowed users to interactively teach a search engine to learn new concepts. Carrie et al. [5] designed an interactive system to judge an input image whether it is a pathological tissue of cancer, and help the pathologist to make a final decision. In order to combine human intelligence with machine capability, Doris et al. [32] proposed a theoretical model to accelerate human-in-the-loop machine learning. Holzinger et al. [17] provided new experimental insights on how people can improve computational intelligence by complementing it with human intelligence in an interactive machine learning approach, and they used the Ant Colony Optimization (ACO) framework to foster multi-agent approaches with human agents in the loop. Wang et al. [30] proposed a crowd-assisted framework, Crowd4ML, which illustrates the corresponding steps to ML. Using the framework, some difficult tasks can be done more clearly. The authors reviewed crowd-assisted machine learning opportunities for future research and identified the main challenges of ML with pure machine intelligence. In 2014, Amerish examined and presented the role of humans in IML and the tasks they can do better [2]. In [28], Sacha et al. proposed a conceptual framework that deals with visual analytic process by identifying key scenarios where ML methods are combined with human feedback through interactive visualization. Justin et al. [9] proposed a hybrid crowd-machine learning classifiers named Flock, it enabled fast prototyping of machine learning models that can improve on both algorithm performance and human judgment, and accomplished tasks where automated feature extraction is not yet feasible. The hybrid systems that use both crowd-nominated and machine-extracted features can outperform those that use either in isolation. Lee et al. [20] proposed a transparent boosting tree (TBT) which visualizes both the model structure and prediction statistics of each step in the learning process of gradient boosting tree to the user and involves user's feedback operations to trees into the learning process. In most of the related works, authors developed a user friendly interface for learning methods and showing significantly improved effects of ML algorithms, gave rise to novel insights of ML models, and integrated both machine capability and human intelligence.

3 THE HUMAN-MACHINE COOPERATION FRAMEWORK

In this section, we first analyze the spatio-temporal feature extraction problem, then we introduce a convolutional autoencoder to detect anomalies. Finally, we introduce human feedback to the video anomaly detection framework.

3.1 Spatio-temporal Feature Extraction

Video sequences contain a large amount of content captured in various contexts. To detect anomalous events or objects of interest, manual examination by watching millions of videos is simply feasible in terms of both time and human resources. On the other hand, using only machines for detection often results in a high error rate, and some samples are difficult to be recognized in complicated surroundings. To address these drawbacks, we develop a novel and practical approach to combing human intelligence with machine computing power. It is a fact that no matter how robust an algorithm is, mis-classifications exist. It is believed that for an anomaly detection framework, people with domain knowledge can give their judgments on the output of the machine, further to avoid retraining the video anomaly detection model. The trained model should reach a respectable precision, such as 0.8. In the test stage, its accuracy on a figure is generally approximately 0.8.

While dealing with the false alarm, we can design a cooperative scheme to address the issue by combining the judgment of experts with the classification result from the machine.

A Convolutional AutoEncoder (CAE) is a feed-forward multi-layer neural network in which the desired output is the input itself. Through several hidden layers of low-dimension features extraction, a non-linear representation of the input data is attained. In particular, autoencoders learn a map from the input to itself through a pair of encoding and decoding phases. The encoder represents features of input frames mapping into hidden layers, and the decoder recovers the hidden representations to the output. The reconstruction error of all pixel values I in frame t of the video sequence is denoted as the Euclidean distance between the input frame feature X_i and the reconstructed frame feature $f_W(X_i)$, represented as Eq. 1:

$$L_{rec} = \frac{1}{N}\sum_{i=1}^{N} ||X_i - f_W(X_i)||_2^2 \tag{1}$$

Considering the above equation is not variable, we add a regularization item to form the objective function as Eq. 2. In order to learn a non-linear classifier, we need to minimize the overall reconstruction errors for the i^{th} training features.

$$L_W = arg\min_{W} \sum_{i=1}^{N} ||X_i - f_W(X_i)||_2^2 + \gamma ||W||_2^2 \tag{2}$$

Where N is the size of the mini batch, γ is a hyper-parameter to balance the loss and regularization and $f_W(\cdot)$ is a nonlinear classifier such as a neural network associated with its weight W [16].

Our framework contains two modules, namely, the video reconstruction error computation and the human-machine cooperation as illustrated in Fig.1.

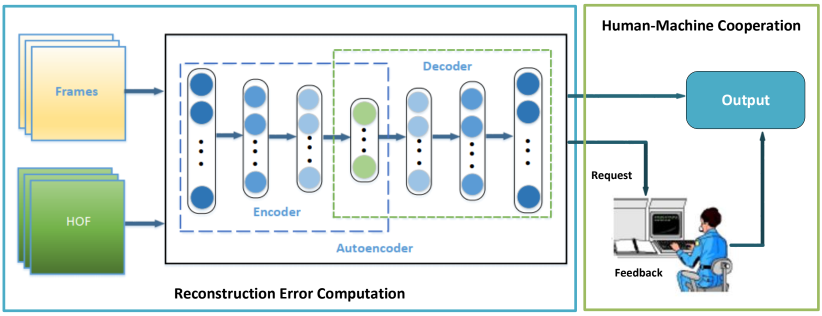


Fig. 1. Human-Machine Cooperation Framework.

3.2 Video Reconstruction Error Computing

We leverage convolutional autoencoder to extract the features of input video frames, then the reconstruction errors are computed by the sum of the small patches divided in each frame. For the normal frames, the reconstruction error is relatively small, while the error is higher for abnormal frames. Once we trained the model, we compute the reconstruction error of a pixel's intensity value I at location (x, y) in frame t of the video sequence. Given the reconstruction errors of the pixels of a frame t, we compute the reconstruction error of a frame by summing up all the pixel-wise errors. A number of frames, such as 10 are fed into the network as a cuboid to obtain the reconstruction error.

3.3 Human-Machine Cooperation Scheme

Due to a certain quantity of false alarms of the CAE based detection, for example, an anomalous object may be largely shielded in a short period, it is inevitable to make a few miss-classifications. To address the problem, we consider incorporating human intelligence into video anomaly detection. At first, there should be several experts to give their judgments and decisions. An expert does not mean the most authoritative and influential person in a certain field but refers to the person who has rich expertise and operation experience, and can well understand and deal with the problems. In our experiments, we used five experts as a decision-making group, and each expert gave his judgment and formed the final decision through voting strategies.

To get a better detection performance, we set three threshold values for the reconstruction error. After the detection of the first stage of the CAE, it has produced a classification of normal and abnormal video frames. Our framework mainly selects the probable abnormal frame according to the reconstruction error which is larger than the threshold. The selected frames will be sent to the human-machine interactive interface for a confirmation or correction. We design an interactive interface to combine the machine detection results with those of human experts. After getting human judgment and correction, the final outputs are highly reliable, which can improve detection accuracy and reduce detection time. We will describe that in more detail in later sections.

4 VIDEO FRAMES RECONSTRUCTION ERROR COMPUTATION

4.1 Convolutional Autoencoder Structure

The structure of the convolutional autoencoder is shown in Fig. 2. The CAE consists of three convolutional layers and two pooling layers in encoder and three deconvolutional layers and two unplooling layers in decoder with a symmetric structure.

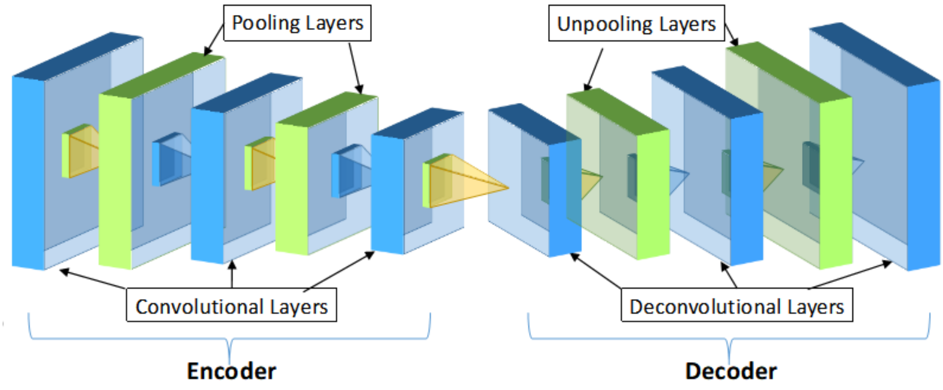


Fig. 2. Convolutional autoencoder structure.

Inspired by the work of [16], we use the CAE to compute the reconstruction errors, with powerful capacity of deep neural network, and high-level features can be learned. In order to obtain the spatio-temporal features of video frames, we stack 10 frames as input to form a cuboid in temporal dimensionality.

Our model resized input frames in 227 × 227 pixels, adapting to the different size of datasets. It has 512 filters with a stride of 4 in the first convolutional layer, thus 512 feature maps with a resolution of 55 × 55 pixels. Both pooling layers have a kernel size of 3 × 3 pixels and perform max pooling. In this step, the patch features are extracted with the spatial dimension compressed output for the following connected layer. Due to the adjacent frames have maximum relevance, the temporal features can be extracted from the consecutive frames and its HOF. The first pooling layer produces 512 feature maps of size 27 × 27 pixels. The second and third convolutional layer have 256 and 128 filters respectively. At last, the encoder produces 128 feature maps of size 13 × 13 pixels. Conversely, the decoder reconstructs the input video frames by deconvolution and unpooling in reverse order of size. The result of the final deconvolutional layer is the reconstructed pattern of the original input.

4.2 Reconstruction Error Computation

In the training process, we observe the accuracy of the CAE model, after thousands of iterations can reach the comparative level of performance to the latest classical work related to the autoencoder. Subsequently, we incorporate human feedback via the interactive interface, later experiments prove it is a practical model to assist the anomaly detection task. While the model is trained, we can validate our model performance on test video frames. To better compare with the work [16], we use a similar formula to calculate the regularity score for all frames, the only difference is the type of learning model. The reconstruction error of all pixel values I in frame t of the video sequence is taken as the Euclidean distance between the input frame and reconstructed frame, shown as Eq. 3, and the un-regularized reconstruction error illustration is shown in Fig. 3.

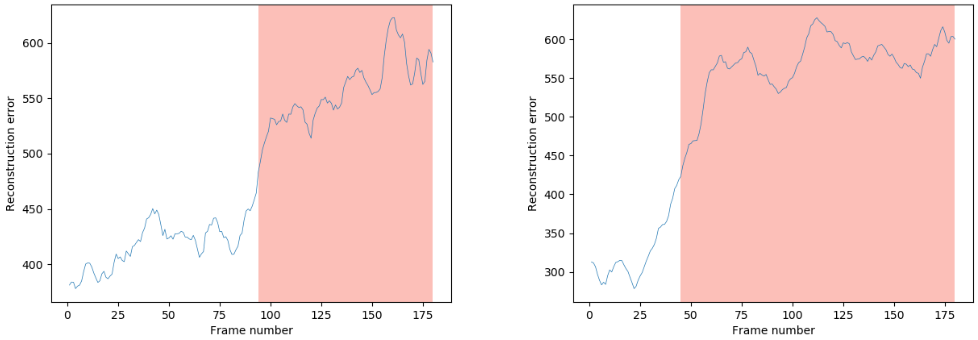


Fig. 3. Un-regularized reconstruction error.

$$e(t) = ||X(t) - f_W(X(t))||_2^2 \tag{3}$$

where f_W is the learned weights by the CAE model. We compute the anomaly score $s_a(t)$ by scaling into the range 0 and 1. Consequently, regularity score $s_r(t)$ can be simply obtained by subtracting anomaly score from 1. Both scores are computed in Eq. 4 and Eq. 5 below, respectively.

$$s_a(t) = \frac{e(t) - e(t)_{min}}{e(t)_{max}} \tag{4}$$

$$s_r(t) = 1 - s_a(t) \tag{5}$$

Video sequences consist of regular events that have a higher regularity score because they are close to the normal training data in the feature space. On the contrary, the anomalous sequences have a lower regularity score, thus it can be used to detect anomalies. Nevertheless, it is impractical to calculate $e(t)_{min}$ and $e(t)_{max}$ in an anomaly detection framework because the future data is un-observable. These two values should be set experimentally according to the historical data [34]. The regularized reconstruction error is illustrated in Fig. 4.

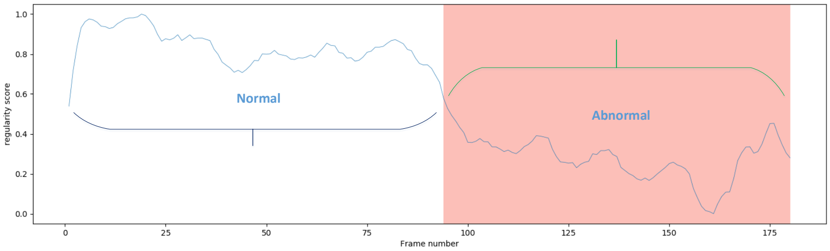


Fig. 4. Regularized reconstruction error.

5 HMCF BASED VIDEO ANOMALY DETECTION

After the training process, the CAE has comparative capacity to address the anomaly detection. However, the classical methods pay more attention to getting a higher precision and lack of analysis of some complex situations. Though the detection precision is gradually improved with time, there are still some intricate situations, so we make analysis and design a human-machine cooperation framework.

5.1 Video Anomaly Scenarios Analysis

Based on previous works [6, 7, 35, 36], we conduct relevant experiments on several datasets, from which we conclude the following three types of difficulties in anomaly detection task.

Heavily occlusions. Under a surveillance camera, the density of moving objects including human and other targets varying from sparse to dense [35], especially, in the case of a group of people walking in a line facing the camera, such as students come out of teaching building after classes, people go to supermarkets on holidays and so on. These situations are common in our daily life, and it is hard to detect the anomaly in such circumstances.

Intense light exposure. On a sunny day, especially at noon, video recordings would attain a sequence of overexposure video frames, this needs a special processing to perform recognition. Besides, in a strong illumination at night, the anomalous objects are still difficult to deal with.

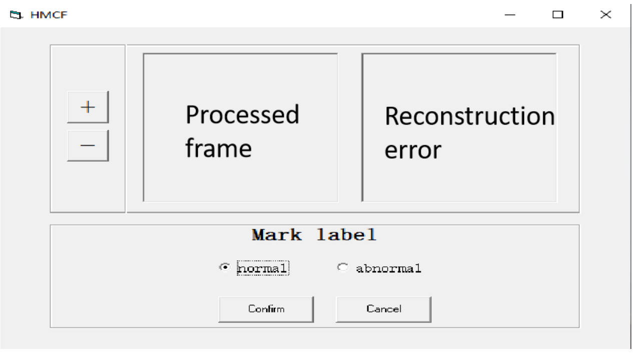
Blurry surveillance video. A shaking camera can generate unclear videos which can impact object detection. The detection performance is variable in terms of the approaches used [6, 7, 36]. It is hard to capture the unambiguous record in extreme weather, such as heavy rain, snowstorm, etc.

5.2 Human Assisted Video Anomaly Detection

We incorporate human feedback into the detection output of CAE, as we mentioned earlier, by setting a proper threshold η, if the reconstruction error of a frame $e(t)$ is larger than η, the frame would be classified as an abnormal, otherwise, it is a normal frame. The output of each frame from the CAE model is denoted by $f(t)$ as shown in Eq. 6.

$$f(t) = \begin{cases} abnormal, & e(t) > \eta \\ normal, & e(t) \leq \eta \end{cases} \tag{6}$$

Our experiments have produced false alarms and missed anomalies. We have analyzed the intricate situations which lead to the above misclassifications, such as occlusions and some regions under the tree in our dataset. We take feedback from experts (having extensive domain knowledge) as the final output. We set requesting frequency for feedback r according to the test video frames N, and the mini-batch number is set to 16 in our experiment. We set the frequency $r = 0.05 * N$. In this step, the ratio can be adjusted according to the detection effect, experts can see the processed video frame as well as its reconstruction error curve of belonged folder as shown in the demo interface Fig. 5. The frames need an expert to make a judgment, these requested frames are stored in a buffer that can be addressed immediately or later, considering that experts do not have to be around machine all the time.



Mark label

● normal ○ abnormal

Fig. 5. Interactive interface for human feedback.

Classification description. There are four types of request cases: i) The frame is classified as an anomaly by the reconstruction error of the CAE model, and experts give a positive label “1" to the interactive interface which corresponds to abnormal and then it is output as a final result; ii) The frame is recognized as an anomaly by the model, while experts judge it as a false alarm, i.e., give the negative label "0" to the option selected box in the display interface. Finally, the frame is detected as normal; iii) The frame is classified as a normal frame by detection model, and experts confirm it is a true normal frame without any abnormal objects, so the final output is normal; iv) The last branch is certainly that a frame is judged as normal by the model, however, it is an abnormal frame indeed judged by experts and mark the frame abnormal as the final output.

Voting strategy. We set up 5 experts to make confirmation for the same requested video frame from CAE. Different expert have his own judgment, in some occasions, they obtain a final identification by majority voting scheme (i.e. for a video frame, if three experts give an “abnormal" label, the final label is “abnormal"). In our datasets, all of the video frames are ordinary scenes, in a campus or a subway scene rather than specific areas such as medical imaging, so it is not hard to judge and confirm. If we research on some special areas, the experts should also have corresponding expertise.

Frame-level anomaly detection. We compute the average reconstruction error of each frame by summing up pixel-wise errors, noted as Per Frame average Error (PFE), and normalize it from 0 to 1. If a PFE is larger than threshold η, its reconstruction error curve fluctuates significantly in the entire video sequence, the curve fluctuation tends to be stable vice versa, as shown in the fig. 4.

Interactive interface. We designed an interface to show detection results as well as the options experts can set. In addition, the local area of a frame can be zoomed in or out in a fixed window in order to scrutinize abnormal regions at fine-grained level. Once a wrong classification is discovered, it can be altered by the option button (normal or abnormal), consequently, the final output frame gets a corrected classification. The interactive interface is shown in Fig. 5.

6 EXPERIMENTS AND EVALUATION

We trained the model using thousands o video frames from multiple video datasets, and evaluated the developed methods with extensive experiments. We run the implemented algorithms and anomaly detection on a HP Server with 4 NVIDIA GeForce GTX 1080 Ti GPUs and 96G memory.

6.1 Datasets

All the video datasets used in this work are publicly available. Each dataset is composed of a number of video clips previously labeled by their creator as normal, or with anomalies. These datasets are widely used in video anomaly detection fields, easy to download, and they can also serve as benchmark datasets for comparison, as these datasets are common scenarios in university or urban, rather than special videos, such as medical imaging, our framework can be extended to many common scenarios.

The UCSD dataset [23] contains two parts, namely UCSD Pedestrian 1 (Ped1) dataset and UCSD Pedestrian 2 (Ped2) dataset. The Ped1 includes 34 training videos and 36 testing ones with 40 irregular events. All of these anomalies include bikers, cars, small trucks and skateboarders and wheelchairs. The Ped2 contains 16 training videos and 12 testing videos with 12 abnormal events. The definition of anomaly for Ped2 is the same as Ped1.

The CUHK Avenue dataset [22] contains 16 training videos and 21 testing videos with a total of 47 abnormal events, including throwing objects, loitering and running. Each clip is about 1 minutes long with a resolution of 640 × 360, having frames range from 50 to 1200.

Subway exit dataset [1] is 43 minutes long with 19 unusual events of two main types: people moving in a wrong direction, people running in the subway platform, loitering near the exit gate. The image resolution is 512 × 384 pixels.

6.2 Experiment Setup

To evaluate our framework, we conducted experiments on three datasets. We trained the CAE at an acceptable level (the detection precision reached about 0.8). Experts are then assigned to give feedback to the requested video frames. They give each frame a final judgment, either a correct classification or a wrong one to be modified. We evaluated HMCF in four aspects: CAE detection performance, effectiveness of anomaly visualization, HMCF operation demonstration, and comparison with different anomaly detection methods.

6.3 CAE Detection Performance

It is straightforward to determine whether a video frame is normal or abnormal. The reconstruction error of each frame determines whether the frame is classified as anomalous. The threshold determines how sensitive we wish the detection approach to behave, for example, setting a low threshold makes the detection become sensitive to the happenings in the scene, where more false alarms would be triggered. We obtain the true positive and false positive rate by setting at different error threshold in order to calculate the area under the receiver operating characteristic (ROC) curve (AUC). The equal error rate (EER) is obtained when false positive rate equals to the false negative rate. We compare the event count with other approaches, which is a significant metrics to show the detection accuracy. In addition, we present the run-time during testing.

Event count. We give the anomalous event count comparison for UCSD dataset, Avenue dataset, and Subway Exit dataset, which is shown in Table 1. For both of UCSD pedestrians scenes, we obtained a comparative level respect to ConvAE [16]. For the Subway Exit scenes, we gained 18 abnormal events compare to ConvAE [16] but at the expense of higher false alarm rate. For Avenue and Subway Exit datasets, we detected more anomalous events and lower false alarms compared to Chong [10]. Our CAE model outperforms or performs comparably to the state-of-the-art abnormal event detection methods but with a few more false alarms. It is because our approach identifies any deviations from regularity, many of which have not been annotated as abnormal events in those datasets.

Table 1. Anomalous event and false alarm count detected by different methods. GT denotes ground truth values of event count.

Method	Anomalous Event Detected / False Alarm			
	UCSD Ped1 (GT:40)	UCSD Ped2 (GT:12)	Avenue (GT:47)	Subway Exit (GT:19)
Chong[10]	N/A	N/A	44/6	18/10
ConvAE[16]	38/6	12/1	45/4	17/5
CAE	37/8	12/3	45/6	18/8

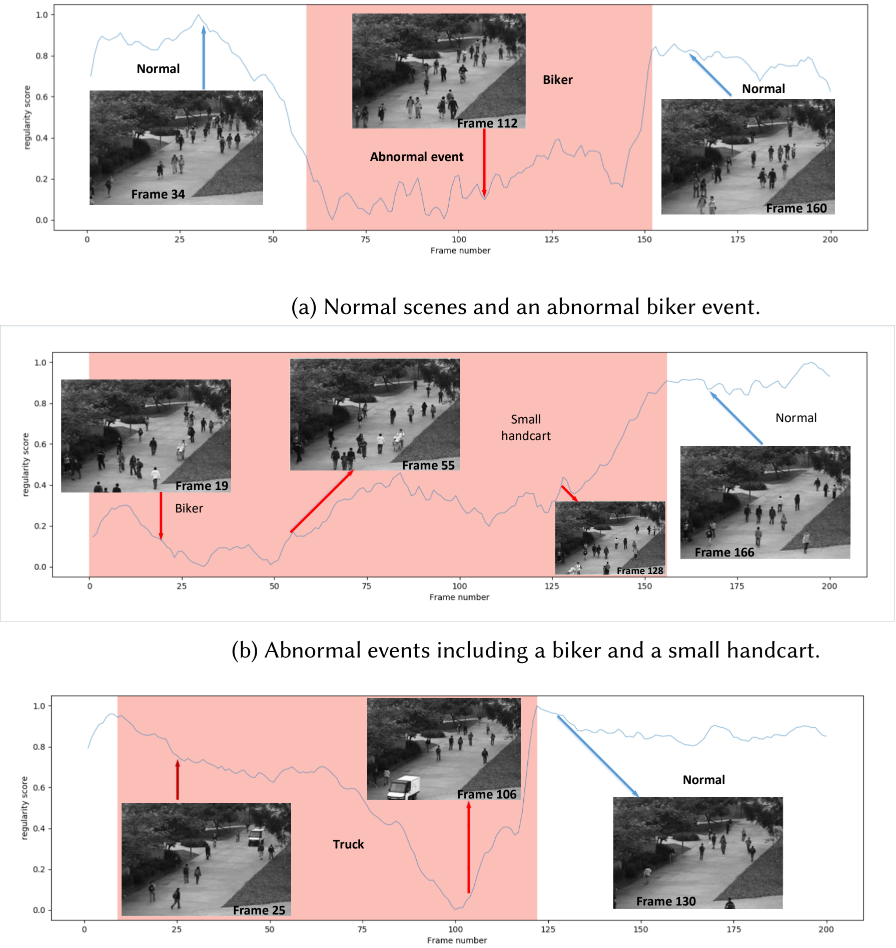
Detection time. We also present an average run-time analysis on CAE event detection, on CPU (Intel Core i5-4590 CPU @3.30GHz) and GPU (NVIDIA GeForce GTX 1080 Ti) respectively in table 2. The total time taken is well less than a quarter second per frame for both CPU and GPU configuration. Due to computational intensive multiplication operations when feeding the input video frames through the convolutional autoencoders, it is wise to run on GPU for a better speed of nearly 30 times faster than CPU.

Table 2. Details of run-time during testing (second/frame)

	Time (Sec)			
	Preprocessing	Representation	Classifying	Total
CPU	0.0010	0.2013	0.0003	0.2026(~5fps)
GPU	0.0010	0.0056	0.0003	0.0069(~145fps)

6.4 Effectiveness of Anomaly Visualization

We obtained the visualizing output of the proposed framework on samples of the Ped1, Ped2, Avenue and Subway Exit dataset, which can detect anomalies correctly in most scenes. Fig. 6 displays normal scenes and abnormal events, such as biker in Fig. 6a and Fig. 6b, a small handcart in Fig. 6b, and a truck in Fig. 6c on Ped1 dataset.



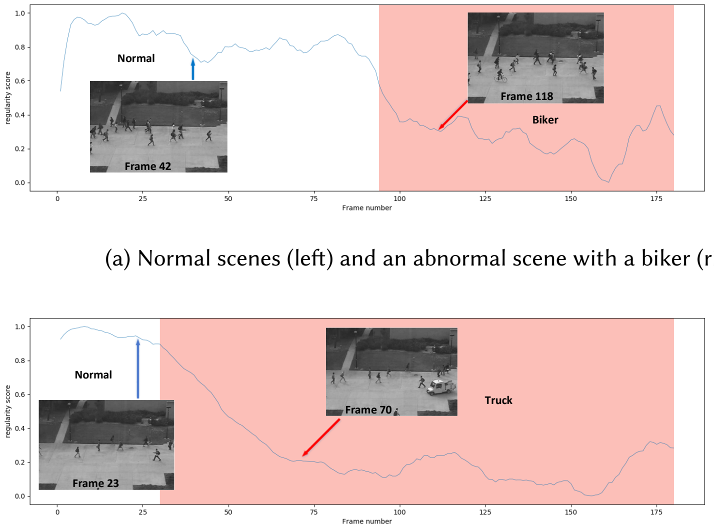
(a) Normal scenes and an abnormal biker event.

(b) Abnormal events including a biker and a small handcart.

(c) Abnormal event while a truck appeared.

Fig. 6. Regularity score of video folders #1, #13 and #27 (from top to bottom) from UCSD Ped1 dataset.

As shown in Fig. 7, anomalies can be detected when an abnormal event appears in a scene, such as a person riding a bicycle in Fig.7a, and a truck in Fig.7b appearing in the surveillance area. In both cases the regularity level show downward trend indicating a low regularity score. Nevertheless, there are situations where anomalies are difficult to be detected. For instance, a skateboarder is easily recognized as an ordinary person walking in a specific view. For some of the miss-classified frames, we can modify them through the interactive interface. We also find that a few abnormal frames are detected as normal, due to the complexity of the objects and events in the scenes. For example, when a group of people walk together, some of them can be occluded by others from the video shooting angle. For the detection algorithm, it could be the case that it works on a number of frames where local areas with important signs are blocked, thus leading to misdetection. This is the original intention of developing this new detection method to incorporate human experience into the detection process.

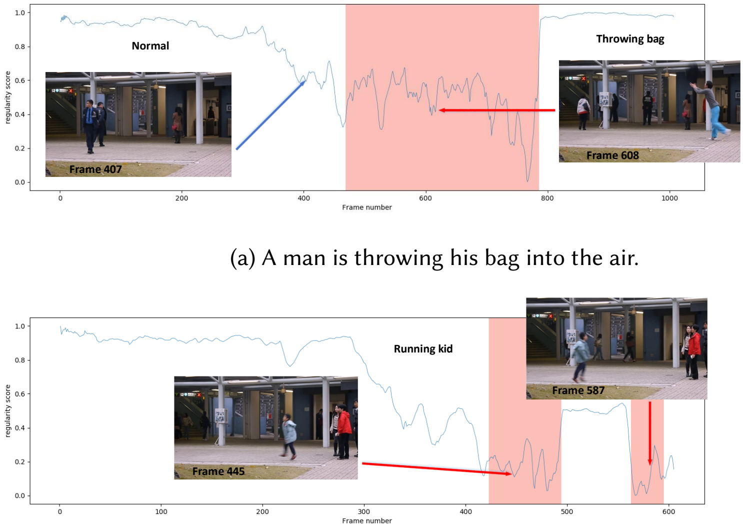


(a) Normal scenes (left) and an abnormal scene with a biker (right).

(b) Abnormal event - a truck appeared in the surveillance area.

Fig. 7. Regularity score of video folders #2, #4 from UCSD Ped2 dataset.

For the Avenue dataset, the detected anomalies are illustrated in Fig. 8, as shown in Fig. 8a, a man is throwing his big into the air. A kid appears in Fig. 8b, he is running back and forth.



(a) A man is throwing his bag into the air.

(b) A kid is running back and forth.

Fig. 8. Regularity score of video folders #5, #7 from Avenue dataset.

For the Subway Exit dataset, we detected abnormal events including running, wrong direction, train approaching and walking group, as shown in Fig. 9, different anomalies are detected and illustrated.

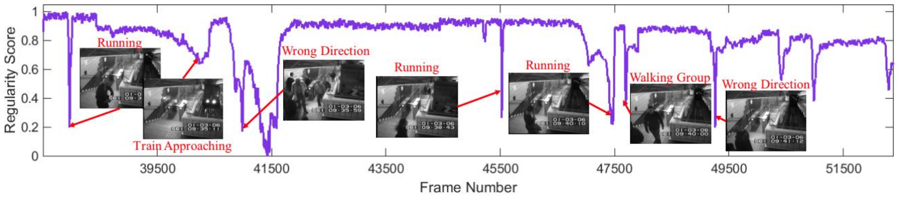
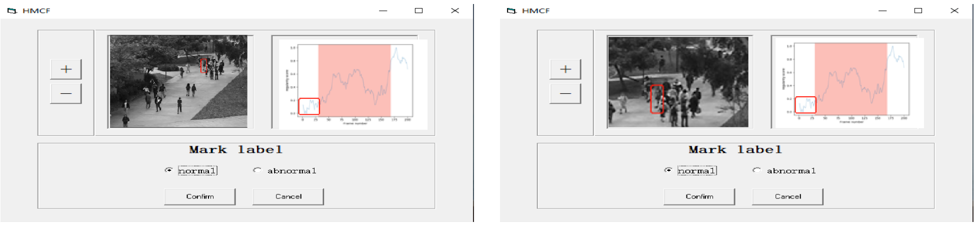


Fig. 9. Abnormal events in the Subway Exit dataset, including running, wrong direction, train approaching and walking group.

6.5 HMCF Operation Demonstration

In the cooperation stage, a certain number of video frames will request human feedback. When an expert finds some mis-classifications, they can modify incorrect classifications produced by CAE in the interface intuitively and select the normal or abnormal option button to confirm the true result. In contrast, when video frames are classified correctly, experts only need to click the “confirm" button to finish the feedback. Fig. 10a shows the skateboarder shielded by a group of people, when just appearing in the range of camera. As a result, it is detected as normal in the CAE model,and the reconstruction error curve is marked with a red rectangle for a short period. For the sake of correctness, the frame can be magnified by clicking the zoom in (“+”) button, consequently, it is clear to confirm an anomaly in the Fig. 10b.



(a) Heavily shielded target

(b) Human correction

Fig. 10. Human judgment.

6.6 Comparison with Different Anomaly Detection Methods

According to the reconstruction error, the regularity score is calculated by Eq. 5 and can be further used to detect anomalous events. As shown in Fig.10, the regularity score of a video clip descends when an anomaly occurs. The salmon color regions denote the ground truth range of anomalous events. Table 3 shows the frame-level AUC and EER performance between the developed method in this paper and other methods with datasets. The AUC improved by 3% ascribe to false alarms and missed anomalies are detected via the framework. In some special situations, the missed anomalies should be paid more attention. Conversely, the EER is decreasing with a lower value. The results show that our approach outperforms all other methods in respect to frame-level AUC, which obtains 93.5% detection accuracy with UCSD Ped1, 93.7% with UCSD Ped2, 83.2% with Avenue dataset, and 94.2% with Subway Exit dataset. Nevertheless, the frame-level EER did not show improvment compare to the approach of Chong [10]. In the CAE model, the objects that not appeared in the train dataset would be detected as an anomaly in the test dataset, however, some of the abnormal events are not marked as anomalies in the ground-truth.

Table 3. Comparison of Area Under ROC Curve (AUC) and Equal Error Rate (EER) of other methods

	Ped1 %		Ped2 %		Avenue %		Subway Exit %	
Method	AUC	EER	AUC	EER	AUC	EER	AUC	EER
Adam[1]	77.1	38.0	-	42.0	-	-	-	-
SF[24]	67.5	31.0	55.6	42.0	-	-	-	-
MPPCA[23]	66.8	40.0	69.3	42.0	-	-	-	-
MDT[23]	81.8	25.0	82.9	25.0	-	-	-	-
ConvAE[16]	81	27.9	90.0	21.7	70.2	25.1	80.7	9.9
STAE[34]	87.1	18.3	88.6	20.9	80.9	24.4	-	-
Chong[10]	89.9	**12.5**	87.4	**12.0**	80.3	20.7	94	**9.5**
CAE	90.1	21.6	90.6	21.3	80.4	23.7	88.6	15.7
HMCF	**93.5**	17.4	**93.7**	18.8	**83.2**	**20.2**	**94.2**	12.6

6.7 Discussion

In our framework, the detection accuracy has prompted by about 3%, especially, in those intricate situations people can show high level recognition of the video frames. The experiments show that our approach is effective in relatively common even crowded scenes. However, there is still great challenge in heavily occlusions or distorted objects detection, human would do little as well as the state of the art methods. Through HMCF, we can build cooperative interface that are both effective and practical. As opposed to off-the-shelf detection algorithms that use thousands of features and require significant effort to understand, HMCF uses the detection results based on video anomaly detection approach, and incorporates human intelligence for intricate scenes. We identify several promising future directions: in improving the feature extraction process, in considering adaptive request frequency, and in devising approach to better automate video features. The design of the HMCF is modular, each component (e.g., reconstruction error computation, human-machine cooperation scheme) can be easily replaced as more effective approaches are found. To make our approach more robust, we can improve HMCF in the following aspects.

Improving feature extraction. We leverage the CAE model to compute the reconstruction error, it is a classical approach for anomaly detection. In fact, there are many other competitive methods, for example, Variational Autoencoder, Gaussian Mixture Model (GMM), fully Convolutional Network (FCN) and so on. Wang et al. [30] proposed a crowd-assisted framework, Crowd4ML, which combines crowd intelligence with deep learning to discover better features. In fact, the central idea of HMCF is the same as the Crowd4ML framework. Our framework currently detects the frame-level anomaly, whereas pixel-wise video anomaly detection will be done in our future work.

Setting the adaptive request frequency. In our experiments, setting the request frequency is a complex problem. If the number is larger, it would increase the burden of experts. On the other hand, if the number is too small, it would not obtain a meaningful result. We set a fixed ratio of request 0.05 according to the experimental performance. Although the number is relatively small for experts, it is not reasonable for a global perspective. It should be adaptive according to the test video frames and reconstruction errors threshold.

Designing interactive interface with more functions. Currently experts can see individual video frames and its corresponding curve, make the final judgment and mark it with the true label. The interactive interface should contain more practical functions, (e.g. capturing the frame image, clipping the image). Once a partly shielded target is clipped, we can store it and further extract its features by the CAE. After iterations, detection performance would be enhanced.

Considering user privacy in video anomaly detection. At present, it is much convenient to obtain surveillance videos from monitoring devices. However, the privacy of users in the videos is not well-considered, and effective measures are lacking to protect some potentially sensitive objects. User privacy has also become a hot spot and challenge in current research. It is a significant direction to encrypt the data set without affecting the effect of anomaly detection. To increase the protection of data, some scholars put an invisibility cloak on the picture, which can well protect the privacy of users, meanwhile people cannot see any changes.

Some thoughts in practical application. The public datasets used in our experiments, come from a single view angle. In real life, there are usually multiple cameras monitoring an area, which can better capture the target objects and abnormal behaviors from multiple perspectives. In practical applications, an intelligent detection algorithm can integrate human judgments into complex detection situations by cooperating with security personnel or related personnel. In the case of severe occlusion, it is not difficult to identify for people, so it can be directly marked up to avoid training model repeatedly.

7 CONCLUSIONS

In this paper, we have developed a novel deep learning based approach for video anomaly detection. Central to the approach is the human-machine cooperation framework is based on multiple CAEs for learning both appearance and motion representations of objects in video frames. The framework can combine learned feature representations with human feedback, thus allowing human experts to help improve anomaly detection performance. We carry out extensive experiments on three public video datasets. The results prove the effectiveness and robustness of the proposed approach, showing the competitive performance with respect to some state-of-the-art methods. Future research will focus on issues such as how to reduce the computation cost, how to incorporate the human feedback in an efficient manner.

ACKNOWLEDGMENTS

This work was supported in part by the National Science Fund for Distinguished Young Scholars (No.61725205), the National Key R&D Program of China (No.2019YFB2102200), and the National Natural Science Foundation of China (No.61960206008).

REFERENCES

[1] Amit Adam, Ehud Rivlin, Ilan Shimshoni, and Daviv Reinitz. 2008. Robust real-time unusual event detection using multiple fixed-location monitors. *IEEE transactions on pattern analysis and machine intelligence* 30, 3 (2008), 555–560. https://doi.org/10.1109/TPAMI.2007.70825

[2] Saleema Amershi, Maya Cakmak, W. Bradley Knox, and Todd Kulesza. 2014. Power to the People: The Role of Humans in Interactive Machine Learning. *AI Magazine* 35, 4 (2014), 105–120. https://doi.org/10.1609/aimag.v35i4.2513

[3] Saleema Amershi, James Fogarty, and Daniel Weld. 2012. ReGroup: Interactive Machine Learning for On-Demand Group Creation in Social Networks. In *Proceeding CHI '12 Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 21–30. https://doi.org/10.1145/2207676.2207680

[4] Y. Benezeth, P. Jodoin, V. Saligrama, and C. Rosenberger. 2009. Abnormal events detection based on spatio-temporal co-occurences. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 2458–2465. https://doi.org/10.

1109/CVPR.2009.5206686

[5] Carrie J Cai, Emily Reif, Narayan Hegde, Jason Hipp, Been Kim, Daniel Smilkov, Martin Wattenberg, Fernanda Viegas, Greg S Corrado, Martin C Stumpe, et al. 2019. Human-centered tools for coping with imperfect algorithms during medical decision-making. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 1–14. https://doi.org/10.1145/3290605.3300234

[6] Karanbir Singh Chahal and Kuntal Dey. 2018. A Survey of Modern Object Detection Literature using Deep Learning. *ArXiv* abs/1808.07256 (2018). arXiv:1808.07256 http://arxiv.org/abs/1808.07256

[7] Rima Chaker, Zaher Al Aghbari, and Imran N. Junejo. 2017. Social network model for crowd anomaly detection and localization. *Pattern Recognition* 61 (2017), 266–281. https://doi.org/10.1016/j.patcog.2016.06.016

[8] Varun Chandola, Arindam Banerjee, and Vipin Kumar. 2009. Anomaly detection: A survey. *ACM Comput. Surv.* 41, 3 (2009), 15:1–15:58. https://doi.org/10.1145/1541880.1541882

[9] Justin Cheng and Michael S Bernstein. 2015. Flock: Hybrid crowd-machine learning classifiers. In *Proceedings of the 18th ACM conference on computer supported cooperative work & social computing*. New York, NY, USA, 600–611. https://doi.org/10.1145/2675133.2675214

[10] Yong Shean Chong and Yong Haur Tay. 2017. Abnormal event detection in videos using spatiotemporal autoencoder. In *International Symposium on Neural Networks*. Springer, 189–196. https://doi.org/10.1007/978-3-319-59081-3_23

[11] Rensso Victor Hugo Mora Colque, Carlos Caetano, Matheus Toledo Lustosa de Andrade, and William Robson Schwartz. 2017. Histograms of Optical Flow Orientation and Magnitude and Entropy to Detect Anomalous Events in Videos. *IEEE Trans. Circuits Syst. Video Techn.* 27, 3 (2017), 673–682. https://doi.org/10.1109/TCSVT.2016.2637778

[12] Yang Cong, Junsong Yuan, and Ji Liu. 2011. Sparse Reconstruction Cost for Abnormal Event Detection. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE Computer Society, USA, 3449–3456. https://doi.org/10.1109/CVPR.2011.5995434

[13] Navneet Dalal and Bill Triggs. 2005. Histograms of oriented gradients for human detection. In *IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, Vol. 1. IEEE, 886–893. https://doi.org/10.1109/CVPR.2005.177

[14] Jerry Alan Fails and Dan R Olsen Jr. 2003. Interactive machine learning. In *Proceedings of the 8th international conference on Intelligent user interfaces*. ACM, 39–45. https://doi.org/10.1145/604045.604056

[15] James Fogarty, Desney Tan, Ashish Kapoor, and Simon Winder. 2008. CueFlik: Interactive Concept Learning in Image Search. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Florence, Italy) *(CHI '08)*. Association for Computing Machinery, New York, NY, USA, 29–38. https://doi.org/10.1145/1357054.1357061

[16] Mahmudul Hasan, Jonghyun Choi, Jan Neumann, Amit K Roy-Chowdhury, and Larry S Davis. 2016. Learning temporal regularity in video sequences. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 733–742. https://doi.org/10.1109/CVPR.2016.86

[17] Andreas Holzinger, Markus Plass, Michael Kickmeier-Rust, Katharina Holzinger, Gloria Cerasela Crişan, Camelia-M Pintea, and Vasile Palade. 2019. Interactive machine learning: experimental evidence for the human in the algorithmic loop. *Applied Intelligence* 49, 7 (2019), 2401–2414. https://doi.org/10.1007/s10489-018-1361-5

[18] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 2012. 3D convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence* 35, 1 (2012), 221–231. https://doi.org/10.1109/TPAMI.2012.59

[19] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*. 1097–1105. https://doi.org/10.1145/3065386

[20] Teng Lee, James Johnson, and Steve Cheng. 2016. An Interactive Machine Learning Framework. *arXiv preprint arXiv:1610.05463* abs/1610.05463 (2016). arXiv:1610.05463 http://arxiv.org/abs/1610.05463

[21] Weixin Li, Vijay Mahadevan, and Nuno Vasconcelos. 2013. Anomaly detection and localization in crowded scenes. *IEEE transactions on pattern analysis and machine intelligence* 36, 1 (2013), 18–32.

[22] Cewu Lu, Jianping Shi, and Jiaya Jia. 2013. Abnormal Event Detection at 150 FPS in MATLAB. In *Proceedings of the 2013 IEEE International Conference on Computer Vision (ICCV '13)*. IEEE Computer Society, USA, 2720–2727. https://doi.org/10.1109/ICCV.2013.338

[23] Vijay Mahadevan, Weixin Li, Viral Bhalodia, and Nuno Vasconcelos. 2010. Anomaly detection in crowded scenes. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, 1975–1981. https://doi.org/10.1109/CVPR.2010.5539872

[24] Ramin Mehran, Alexis Oyama, and Mubarak Shah. 2009. Abnormal crowd behavior detection using social force model. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 935–942. https://doi.org/10.1109/CVPR.2009.5206641

[25] Vignesh Ramanathan, Jonathan Huang, Sami Abu-El-Haija, Alexander Gorban, Kevin Murphy, and Li Fei-Fei. 2016. Detecting events and key actors in multi-person videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3043–3053. https://doi.org/10.1109/CVPR.2016.332

[26] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2017. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39, 6 (2017), 1137–1149. https://doi.org/10.1109/TPAMI.2016.2577031

[27] Mohammad Sabokrou, Mahmood Fathy, Mojtaba Hoseini, and Reinhard Klette. 2015. Real-time anomaly detection and localization in crowded scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. 56–62. https://doi.org/10.1109/CVPRW.2015.7301284

[28] Dominik Sacha, Michael Sedlmair, Leishi Zhang, John Lee, Jaakko Peltonen, Daniel Weiskopf, Stephen North, and Daniel Keim. 2017. What You See Is What You Can Change: Human-Centered Machine Learning By Interactive Visualization. *Neuro computing* 268 (04 2017). https://doi.org/10.1016/j.neucom.2017.01.105

[29] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. 2008. Extracting and Composing Robust Features with Denoising Autoencoders. In *Proceedings of the 25th International Conference on Machine Learning* (Helsinki, Finland) *(ICML '08)*. ACM, New York, NY, USA, 1096–1103. https://doi.org/10.1145/1390156.1390294

[30] J. Wang, Y. Wang, and Q. Lv. 2019. Crowd-Assisted Machine Learning: Current Issues and Future Directions. *Computer* 52, 1 (2019), 46–53. https://doi.org/10.1109/MC.2018.2890174

[31] Shandong Wu, Brian E Moore, and Mubarak Shah. 2010. Chaotic invariants of lagrangian particle trajectories for anomaly detection in crowded scenes. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, 2054–2060. https://doi.org/10.1109/CVPR.2010.5539882

[32] Doris Xin, Litian Ma, Jialin Liu, Stephen Macke, Shuchen Song, and Aditya G. Parameswaran. 2018. Accelerating Human-in-the-loop Machine Learning: Challenges and Opportunities. In *Proceedings of the Second Workshop on Data Management for End-To-End Machine Learning, DEEM@SIGMOD 2018, Houston, TX, USA, June 15, 2018*. 9:1–9:4. https://doi.org/10.1145/3209889.3209897

[33] Dan Xu, Yan Yan, Elisa Ricci, and Nicu Sebe. 2017. Detecting Anomalous Events in Videos by Learning Deep Representations of Appearance and Motion. *Computer Vision and Image Understanding* 156 (March 2017), 117–127. https://doi.org/10.1016/j.cviu.2016.10.010

[34] Yiru Zhao, Bing Deng, Chen Shen, Yao Liu, Hongtao Lu, and Xian-Sheng Hua. 2017. Spatio-temporal autoencoder for video anomaly detection. In *Proceedings of the 25th ACM international conference on Multimedia*. ACM, 1933–1941. https://doi.org/10.1145/3123266.3123451

[35] Zhong-Qiu Zhao, Peng Zheng, Shou-tao Xu, and Xindong Wu. 2019. Object Detection With Deep Learning: A Review. *IEEE Trans. Neural Netw. Learning Syst.* 30, 11 (2019), 3212–3232. https://doi.org/10.1109/TNNLS.2018.2876865

[36] Chong Zhou and Randy C. Paffenroth. 2017. Anomaly Detection with Robust Deep Autoencoders. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Halifax, NS, Canada, August 13 - 17, 2017*. 665–674. https://doi.org/10.1145/3097983.3098052

Received June 2020; revised June 2020; accepted July 2020