A Fine Grained Quality Assessment of Video Anomaly Detection

Jiang Zhou
Dublin City University
Dublin, Ireland
jiang.zhou@dcu.ie

Kevin McGuinness
Dublin City University
Dublin, Ireland
kevin.mcguinness@dcu.ie

Joseph Antony
Dublin City University
Dublin, Ireland
josejaihind@gmail.com

Noel E. O'Connor
Dublin City University
Dublin, Ireland
noel.oconnor@dcu.ie

ABSTRACT

In this paper we propose a new approach to assess the performance of video anomaly detection algorithms. Inspired by the COCO metrics we propose a quartile based quality assessment of video anomaly detection to have a detailed breakdown of algorithm performance. The proposed assessment divides the detection into five categories based on the measurement quartiles of the position, scale and motion magnitude of anomalies. A weighted precision is introduced in the average precision calculation such that the frame-level average precision reported in categories can be compared to each other regardless of the baseline of the precision-recall curve in every category.

We evaluated three video anomaly detection approaches, including supervised and unsupervised approaches, on five public datasets using the proposed approach. Our evaluation shows that the anomaly scale introduces performance difference in detection. For both supervised and unsupervised methods evaluated, the detection achieve higher average precision for the large anomalies in scale. Our assessment also shows that the supervised multiple instance learning method is robust to the motion magnitude differences in anomalies, while the unsupervised one-class neural network method performs better than the unsupervised autoencoder reconstruction method when the motion magnitudes are small. Our experiments, however, also show that the positions of the anomalies have impact on the performance of the multiple instance learning method and the one-class neural network method but the impact on the autoencoder-based approach is negligible.

CCS CONCEPTS

• **Computing methodologies** → *Scene anomaly detection.*

KEYWORDS

video anomaly detection, neural networks, evaluation metrics

CBMI 2022, September 14–16, 2022, Graz, Austria

ACM ISBN 978-1-4503-9720-9/22/09.
https://doi.org/10.1145/3549555.3549569

ACM Reference Format:
Jiang Zhou, Kevin McGuinness, Joseph Antony, and Noel E. O'Connor. 2022. A Fine Grained Quality Assessment of Video Anomaly Detection. In *International Conference on Content-based Multimedia Indexing (CBMI 2022), September 14–16, 2022, Graz, Austria.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3549555.3549569

1 INTRODUCTION

Detecting abnormal events is critical in video surveillance and is traditionally labor intensive. In recent years computer vision algorithms have been developed in an effort to automatically detect the abnormal events from videos. The performance of these algorithms has been measured by the area under receiver operating characteristic curve (auROC), equal error rate (EER), F1 score, etc. These metrics give an assessment of the overall performance of each method. The diverse nature of anomalies, however, is a big challenge to their detection, and fine-grained assessment is needed to establish the advantages and disadvantages of various algorithms.

The COCO metrics are widely used in object detection [10]. Rather than computing an average precision (AP) for all objects, COCO metrics compute APs across multiple object scales. Inspired by the COCO metrics, we introduce a detailed breakdown analysis of the video anomaly detection in terms of the position, scale and motion magnitude of the anomalies. However, the principle behind the thresholds for defining scale categories in COCO metrics is unclear. In this paper we propose using the quartiles as the basis to break down the measurement. Quartiles provide information about both the center and the spread of the data, and statistics based on quartiles such as the interquartile range (IQR) are not unduly affected by outliers. With the quartiles, we divide the anomalies into five categories and name them as "tiny," "small," "medium," "large," and "huge," and the frame-level interpolated average precision is calculated in every category. The AP computation is, however, strongly impacted by the number of anomalies in each category. We therefore propose a weighted precision in the AP calculation such that the reported APs are comparable to each other.

With the proposed assessment, we evaluate three video anomaly detection methods. The first method applies the one-class neural networks (OCNN) [2] for video anomaly detection. The method was designed to detect anomalies in images but hasn't been evaluated for detecting anomalies in videos. It represents those one-class models trained with normal videos and detecting anomalies by measuring their deviations from trained models. The second method uses the autoencoder (AE) to learn the regular motion patterns for anomaly

detection [6]. It represents those reconstruction-based approaches that use the reconstruction error as a proxy for anomaly score. Many of these reconstruction-based approaches need per-video score normalization for detection. This is problematic as it assumes there is at least one anomalous event in the video but the end of the test video is unknown in practice. To remove such a limitation we propose applying a 1D derivative filter on the anomaly scores and detecting the anomalies based on the adjacent anomaly score difference. The third method in our evaluation uses the multiple instance learning (MIL) [19] for the detection. It represents those approaches with the idea of weakly supervised training for anomaly detection.

Five public datasets are used in the evaluation. The videos from the UCSD Pedestrian [14], ShanghaiTech [12], and Street Scene [16] datasets are captured in constrained conditions and scenes, while the datasets Live Video [9] and UCF Crime [19] contain CCTV footage collected in the wild. Our evaluation on these datasets shows that the anomaly scale does have impact on the detection. For all three methods evaluated the detection achieve higher average precision for the anomalies large in scale. However, the OCNN method introduces the least variance in performance for different anomaly scales while the AE method is the most affected by the scale differences. Our assessment also shows the positions of the anomalies have impact on the performance of the MIL method and the OCNN method but not on the autoencoder reconstruction method. In the experiment results, the supervised MIL method introduces much less variances in detection than the unsupervised OCNN method for different anomaly positions. For the motion magnitude measurement, the weakly supervised MIL method demonstrates its detection robustness to the differences in the motion magnitudes of anomalies while the OCNN shows a better capacity to handle the variance in motion magnitude than the AE method.

1.1 Video anomaly detection

Video anomaly detection has gained increasing attention in recent years and many papers have demonstrated promising results. Classic approaches usually require substantial feature engineering. Cong et al. [3] extract multi-scale histogram of optical flow to build a normal dictionary for calculating the sparse reconstruction cost of test samples. Lu et al. [13] extract 3D gradient features of video patches at multiple scales in their sparse combination learning for fast abnormal event detection. Doshi et al. [4] extracts motion, location, and appearance features for each object in a frame and learns the normal frames using few-shot learning with k-nearest neighbor (k-NN) distances. If the k-NN distance of any object with respect to a training set is larger than a threshold, the frame with that object is deemed anomalous.

With the advent of deep learning, which presents a way to learn relevant features automatically, a vast majority of recent video anomaly detection methods depend on deep neural networks. As anomalous events usually come as rare cases while videos with normal events are easy to be obtained, many methods pursue the unsupervised learning for the problem without providing anomalies for supervision. Tran et al. [21] use convolutional winner-take-all autoencoders to learn motion-feature representations from optical flow fields of videos and feed learnt sparse representations to a one-class SVM to perform anomaly scoring. However, such hybrid approaches with SVM is suboptimal as they are unable to influence representational learning in the hidden layers [7]. Ruff et al. [18] therefore train a neural network while minimizing the volume of a hypersphere that encloses the network representations of data. The network is then forced to extract the common factors of variation of normal samples when used for anomaly detection tasks. Chalapathy et al. [2] further combine the ability of deep neural networks to extract progressively rich representation of data along with the one-class objective for the anomaly detection.

Hasan et al. [6] train a convolutional autoencoder to reconstruct video snippets for the anomaly detection. It is assumed that the learned autoencoder reconstructs regular motion with low error but incurs high reconstruction error for irregular motions. With a similar motivation, Ravanbakhsh et al. [17] attempt the first use of generative adversarial networks (GANs) to learn the normality of the crowd behaviour. Since the GANs are trained with only normal data, they are unable to reconstruct appearance and motion information of the possible abnormal regions of the test frame. A simple difference between the real test-frame representations and the generated descriptions would detect abnormal areas in the frame. These approaches have, however, been criticized because large reconstruction errors can not be guaranteed for abnormal events [12]. Liu et al. [12] therefore propose to tackle the problem within a video prediction framework. A future frame is predicted with an adversarial trained U-Net. The abnormal frame is then detected by thresholding the normalized peak signal to noise ratio between the predicted frame and the real frame.

It is argued that an anomaly detection system can perform better with the training data of both normal and anomalous events [7, 19]; however, the annotation of anomalous events is very expensive. Sultani et al. [19] therefore propose a weakly supervised learning approach with only video-level annotations. The approach learns anomaly through a deep multiple instance learning (MIL) framework by treating normal and anomalous videos as bags and short segments of each video as instances in a bag. A bag is labelled as positive if any segment in the bag contains anomaly, otherwise the bag is labelled as negative. Following the same path, Kamoona et al. [7] add a temporal encoding-decoding network to the MIL ranking framework to capture the temporal information among the video clips. Dubey et al. [5] propose 3D deep MIL with ResNet along with a new proposed ranking loss to mitigate the false alarm rate in anomaly detection.

For certain scenarios such as road and highway traffic, supervised learning can be designed [1, 22], where the annotations of the anomalies in the training videos are available. However, because it is impossible to list all anomalous events, these approaches fail to generalize well to other applications. Moreover, the supervised approaches defeat the spirit of anomaly detection where the ultimate goal in practice is to detect any deviation from normality [15].

1.2 Frame-level evaluation

The performance of the anomaly detection is commonly measured by the temporal localization of anomaly events, and therefore the frame-level criteria are widely used in the evaluation. The frame-level auROC is usually reported in video anomaly detection. The ROC curve needs 1-specificity in the plot, which takes the proportion of correctly identified normal frames into account. Class

imbalance is, however, usually an issue in anomaly detection where the true negatives are much more common than the true positives; therefore, the auROC is not very suitable for the performance assessment. The F1 score is the harmonic mean of the precision and recall and can be used an alternative metric for the evaluation, but an optimal threshold needs to be determined before the score calculation.

In this paper, average precision (AP) is used as the metric for the performance. AP is the area under the precision-recall (PR) curve, which focuses on predictions around the anomaly class. A detected anomaly frame can be counted as a true positive if this frame has ground truth anomalous regions, otherwise a false positive is counted. To remove the zig-zag pattern in the PR curve, the interpolated average precision [10] is used.

2 DATASETS

Five public datasets are used in our assessment as shown in Figure 1. UCSD Pedestrian [14] is one of the most widely used datasets. It contains two subsets, UCSD Ped1 and Ped2. Each subset contains videos captured from a static camera overlooking a pedestrian walkway. Pedestrian walking is considered as normal and anomalies are presented as bikers and skaters, etc. The Ped1 subset contains 34 normal videos as training data and 36 test videos. The Ped2 subset consists of 16 training videos and 12 test videos. Pixel-level and frame-level annotations are available for both Ped1 and Ped2 test videos. Bounding box annotations of anomalous regions can also be found [16].

(a) UCSD (b) ShanghaiTech (c) Street Scene

(d) Live Video (e) UCF Crime

Figure 1: Five public datasets used in assessment.

ShanghaiTech [12] dataset contains videos from 13 different scenes with complex light conditions and camera angles. Same as UCSD Pedestrian dataset, pedestrian walking is the normal pattern and anomaly events are introduced by cars or bikers, etc. It consists of 330 normal videos and 107 test videos with 130 anomalies. Both frame-level and pixel-level annotations are provided for the test videos.

Street Scene [16] is a recent proposed dataset that contains 46 training and 35 test videos with a total of 205 anomalous events consisting of 17 different anomaly types. The videos are taken from a static USB camera looking down on a scene of a two-lane street with bike lanes and pedestrian sidewalks, hence the anomaly instances are "jaywalking," "biker outside lane," "loitering," etc. Bounding boxes and frame-level annotations are available for the test videos.

It is argued that the datasets mentioned above do not reflect realistic anomalies in video surveillance. Therefore, some datasets are prepared with videos captured in unconstrained, "in-the-wild" conditions and scenes. The Live Video dataset [9] consists of videos from surveillance cameras depicting realistic abnormal events. It contains 30 videos with 14 types of abnormal events such as fighting, people clashing, kidnapping, etc. Bounding boxes in anomalous frames are given in the dataset. The UCF-Crime dataset [19][1] is a recently proposed dataset with realistic CCTV videos. The dataset contains 950 normal videos and 950 unedited videos covering 13 real-world anomaly events such as "abuse," "burglary," "shoplifting," etc. Among the 1900 videos, 150 normal videos and 140 anomalous videos are given frame-level annotations for validation.

3 EVALUATION METHODS

Our selected datasets are evaluated with 3 methods for anomaly detection, the one-class neural networks (OCNN) method, the autoencoder (AE) reconstruction method and the multiple instance learning (MIL) method. All three methods take features extracted from a C3D network [20] pretrained on Sports-1M [8]. C3D is a widely used neural network for extracting generic video features. For each video, we resize every video frame to 112×112 pixels and compute the C3D features for non-overlapping 16-frame clips from the video. The visual features are taken from the fully connected (FC) layer FC6 of the C3D with L_2 normalization. This gives each 16-frame clip a 4096 dimensional descriptor.

3.1 One-class neural network

Approaches using one-class classification for anomaly detection aim to learn the models that can accurately describe the "normality." Following Chalapathy and Chawla [2], a one-class neural network (OCNN) is trained in our evaluation. The network takes a clip descriptor as input and feeds forward to a fully connected layer with 512 neurons. Both the input layer and the hidden layer use 20% dropout. The hidden layer connects to the scalar output of the network, and drives its data representation with the objective:

$$\min_{w,V,r} \frac{1}{2}||w||_2^2 + \frac{1}{2}||V||_F^2 + \frac{1}{\mathcal{V}} \cdot \frac{1}{N}\sum_{n=1}^{N} \max\{0, r - \langle w, g(VX_{n:})\rangle\} - r$$

where V is the weight matrix from input to the hidden units, w is the weights from the hidden layer to the output, and g is the sigmoid activation function. In the objective, $\mathcal{V} \in (0, 1)$ controls the training error allowed, which is set as 0.1 in our experiments. r gives the bias of the hyper-plane that separates normal and anomalous data.

3.2 Autoencoder reconstruction

An autoencoder (AE) is built in our evaluation with the assumption that the anomalies can be detected with high autoencoder reconstruction errors. The autoencoder takes each C3D clip descriptor as input and passes through three hidden layers with 512, 32, and 512 neurons respectively. The autoencoder outputs a reconstructed 4096 dimensional vector and the mean squared error (MSE) loss is measured between the input and the output with 5e-4 weight decay during training. All layers in the autoencoder are fully connected

[1]The bounding box annotation can be downloaded from https://github.com/xuzero/UCFCrime_BoundingBox_Annotation [11]

layers with 20% dropout, and the rectified linear unit (ReLU) is used as the activation function.

During testing many methods apply the per-video normalization [6], which scales the reconstruction error to range $[0, 1]$. However, this assumes that every test sequence has at least one normal and one anomalous frames, and it is impractical for real-world applications because the end of a test video sequence is unknown. Therefore, we propose a novel way to detect the anomalies based on the difference between adjacent frame reconstruction error. Given a test video, a sequence of MSEs is computed from the autoencoder. This sequence of MSEs is then padded with the mean of MSEs from all training samples at the top and the end of the sequence, and is convolved with a 1D kernel [1, 0, -1] to generate a sequence of MSE difference scores. Within the difference scores, a high positive difference score signifies the start of an anomalous event and a low negative score corresponds to the end of the event. Consequently all frames between the pair of the signals are classified as anomalous.

3.3 Multiple instance learning

It is contended that the anomaly detection performance can be improved by introducing supervision information. We therefore follow Sultani et al. [19] and train models with both normal and anomalous data using multiple instance learning (MIL). Each video is divided into 32 temporal segments and a C3D descriptor is extracted from each segment. The 4096 dimensional vector is input to a feed-forward network with two hidden layers before the scalar output anomaly score. Both hidden layers are fully connected as in Sultani et al. [19], but one has 512 neurons and the other has 32 neurons. The MIL ranking loss [19] is used for training.

4 ASSESSMENT

Table 1 shows the AP of the three methods on the five datasets. The baseline of AP for each dataset is also given in the table, which is calculated as P/N. P is the total number of the positive frames in a dataset and N is the total number of frames in the dataset. There is no obvious winner method that can be identified in the table across all 5 datasets, and thus a detailed assessment to characterize every method would be very helpful. For a fine-grained assessment of

datasets	baseline	OCNN	AE	MIL
UCSD	0.68	0.71	0.60	0.73
Shanghai	0.42	0.51	0.58	0.44
Street	0.30	0.32	0.37	0.35
Live	0.37	0.57	0.49	0.70
UCF	0.06	0.06	0.12	0.17

Table 1: The APs of three methods on 5 datasets. UCSD stands for the UCSD Pedestrian dataset. Shanghai stands for the ShanghaiTech dataset. Street is short for Street Scene and Live is short for Live Video. UCF is the UCF Crime dataset.

detection performance, the anomalies in every dataset is measured in terms of their positions, scales and motion magnitudes. For each measurement, two quartiles can be calculated in every dataset, the first quartile Q_1 and the third quartile Q_3. The Q_1 gives the cut point for the lowest 25% of the measured data while the Q_3 splits off the highest 25%. With Q_1 and Q_3, it is possible to calculate the interquartile range (IQR) as $Q_3 - Q_1$, a robust measure of data dispersion. Outliers can then be checked for using the lower fence and the upper fence, where the lower fence (LF) is calculated as $Q_1 - 1.5 \times IQR$ and the upper fence (UF) is calculated as $Q_3 + 1.5 \times IQR$. The cut points Q_1, Q_3, LF and UF divide the anomalies into five categories $(-\infty, LF)$, $[LF, Q1)$, $[Q1, Q3)$, $[Q3, UF)$, and $[UF, +\infty)$ and we name them "tiny," "small," "medium," "large," and "huge" correspondingly.

The quartiles are calculated in every dataset and thus the analysis based on the quartiles are not comparable to each other across datasets. To solve this problem, the average values of the cut points across the five datasets for each measurement are calculated. However, in this way the Q_1 and Q_3 may be no longer the 25th and 75th percentiles of the data. Therefore, the percentage of the anomalies in each category needs to be calculated for every dataset. A precision is calculated as $tp/(tp + fp)$ where tp is the number of the true positives and fp is the number of the false positive. The number of the false positives is unrelated to how categories of anomalies are defined and the precision is strongly impacted by the number of positive samples in each category. Consequently the AP computed in each category is not comparable to each other due to different numbers of anomalies in categories. To tackle this issue, we propose a weighted precision in the AP computation, where the precision is calculated as $tp/(tp + fp \times p_1)$ where p_1 is the percentage of anomalies in each category.

4.1 Position

The position measures how close an anomaly bounding box is to the image borders, and we are interested in whether different positions of the anomalies make any difference in the detection. To be independent of frame resolutions across datasets, the positions are measured as ratios. Given the center of a bounding box x, y and the image width w and height h, the position is calculated as the lowest ratio value among the x/w, y/h, $(w-x)/w$ and $(h-y)/h$. With the proposed quartile based metrics, the average LF, Q_1, Q_3, and UF for the position measurement are $[-0.089, 0.137, 0.288, 0.515]$. Table 2 shows the percentage of anomalous frames in each range part for the 5 datasets. If multiple anomalies appears in the same frame, the largest position is used to categorize the frame to a category. Because the minimum possible value for a position is 0 and the maximum is 0.5, there can not be extremes in the tiny and huge categories as shown in Table 2. Most anomalies in the Street Scene dataset are near the frame borders and most anomalies in the UCF Crime dataset are away from the frame borders. This is because videos from the Street Scene dataset are taken from a static camera looking down on a scene of a street with bike lanes and pedestrian sidewalks. The bike lanes and sidewalks are close to the image borders and most anomalies in the Street Scene dataset are about pedestrians and bikers. The videos from the UCF Crime dataset are

	tiny	small	medium	large	huge
UCSD	0	10.89%	28.96%	60.15%	0
Shanghai	0	31.85%	36.33%	31.83%	0
Street	0	70.48%	18.19%	11.34%	0
Live	0	23.82%	54.67%	21.51%	0
UCF	0	4.35%	26.27%	69.38%	0

Table 2: The percentages of anomalous frames in each category under the position measurement.

selected with clear anomalies; therefore, most anomalous events are near the center of the frames. The anomalies in the ShanghaiTech dataset are pedestrians and vehicles that move across the scene from one end to the other; therefore, the number of anomalies in the small, medium, and large categories are close.

The frame-level AP with weighted precision is calculated in each category. We turn on dropout during inference and run every method 15 times for the five datasets. This gives 75 AP results in each category for a method. However, the 5 datasets have different baselines of AP, which make the analysis across datasets difficult. We therefore subtract the baseline from the APs for each dataset and use the residuals in an one-way analysis of variance (ANOVA). The residuals from the small, medium, and large categories form 3 groups in the ANOVA and each group has 75 AP residuals. The null hypothesis is tested, which states that the means of the AP residuals from the small, medium, and large position categories have no significant differences to each other.

	OCNN	AE	MIL
p value	0.00004	0.315	0.009
f value	10.543	1.160	4.809

Table 3: ANOVA results for the position measurement.

	tiny	small	medium	large	huge
OCNN	0	0.029	0.080	0.067	0
AE	0	0.087	0.071	0.062	0
MIL	0	0.132	0.108	0.079	0

Table 4: The means of AP residuals in each category under the position measurement.

Table 3 shows the ANOVA results. Given the α level 0.05, it can be seen that the prediction from the autoencoder reconstruction method is unimpacted by the positions of the anomalies. This is because the reconstruction error from the AE method is a sum of each feature difference, which is irrelevant to the anomaly position. It is surprising that there are performance differences among the groups for the one-class neural network method and the multiple instance learning method under the position measurement. These differences can also been found in Table 4, which lists the means of the AP residuals in each category. However, the f values in Table 3 show that, as a supervised learning method, MIL introduces much less variance in detection than the unsupervised OCNN method for different positions.

4.2 Scale

The scale measures the relative size of an anomaly in a frame. Given an anomaly, the scale is calculated as the ratio of the size of the bounding box to the size of the frame. As the position measurement, the average LF, $Q1$, $Q3$, and UF of the scale measurement across the 5 datasets are calculated, which are $[-0.069, 0.025, 0.088, 0.183]$. The anomalous frames from all datasets are then categorized into the 5 categories of scale. Table 5 shows the percentage of the anomalous frames in each category. If multiple anomalies appears in the same frame, the largest scale in the frame is used to categorize the frame.

Most of the anomalies from the top 3 datasets in Table 5 are relatively small. The UCSD, ShanghaiTech, and the Street Scene datasets are prepared under constrained conditions; therefore, the variance of the scale is small in these datasets. In contrast, the Live Video dataset and the UCF Crime dataset are prepared in unconstrained, "in-the-wild" conditions, so both datasets show a certain variety of scales. However, the anomalies in the UCF Crime dataset tend to be relatively large. One reason is that a strategy of having all anomalies enclosed within a single large bounding box in a frame was used for the annotation of the UCF Crime dataset.

	tiny	small	medium	large	huge
UCSD	0	78.67%	21.33%	0	0
Shanghai	0	66.00%	28.87%	4.68%	0.44%
Street	0	96.66%	3.22%	0.12%	0
Live	0	27.07%	54.15%	10.81%	7.96%
UCF	0	2.31%	22.17%	20.31%	55.21%

Table 5: The percentages of anomalous frames in each category under the scale measurement.

	OCNN	AE	MIL
p value	0.012	0.0001	0.007
f value	3.740	7.206	4.159

Table 6: ANOVA results for the scale measurement.

	tiny	small	medium	large	huge
OCNN	0	0.079	0.013	0.167	0.146
AE	0	0.046	0.212	0.123	0.169
MIL	0	0.135	0.067	0.159	0.157

Table 7: The means of AP residuals in each category under the scale measurement.

As with position measurement, we run every method 15 times for each dataset and calculate AP residuals in each category as a group in the ANOVA. Table 5 shows the UCSD dataset has no frames with anomalies in the large and huge categories. The Street Scene dataset has only 0.12% anomalous frames in the large category and no anomalies in the huge category while the other three datasets have frames with anomalies in all small, medium, large, and huge categories. Therefore, we take a two-sided t-test first with the small and medium categories, which have AP residuals from all datasets, and calculate p-values of $[0.006, 5.6 \times 10^{P} - 19, 0.004]$ and t-values of $[2.81, -10.25, 2.92]$ for OCNN, AE, and MIL respectively. ANOVA then is run with the 4 categories of AP residuals from the ShanghaiTech, Live Video, and UCF Crime.

Table 6 shows the ANOVA results. It can be seen that all the p-values from both the t-test and the ANOVA are smaller than the α level 0.05, so there are significant differences among group means for the three methods. In other words different scales of anomalies introduce significant difference in the anomaly detection performance. This can also be seen in Table 7. When the anomalies are large or huge, the detection performance is better than the detection in the small and medium categories for the OCNN and MIL methods. AE achieves the best detection in the medium category and the performance in the large and huge categories are much

better than the small category. Moreover, the t-values and f-values show that OCNN introduces the least performance variance among the categories while the AE method is the most affected by the scale differences.

4.3 Motion magnitude

The motion magnitude is calculated as the average magnitude of the optical flow inside an anomaly bounding box. Video frames in all datasets are resized to size 112×112 and smoothed by a Gaussian filter with kernel size 7×7 before computing the optical flow. The Farneback algorithm from OpenCV is used to compute the dense optical flow. The pyramid scale is set to 0.5 and the pyramid levels is set to 3. The algorithm runs 3 iterations at each pyramid level with a window size of 10. The size of the pixel neighborhood for polynomial expansion in the algorithm and the standard deviation for smooth derivatives are set to 5 and 1.2. After computing the motion magnitude for all anomalies, the cutting points of the 5 categories are obtained as $[-0.369, 0.107, 0.424, 0.900]$ for the motion magnitude measurement. Table 8 shows the percentage of anomalous frames in each category. If multiple anomalies appears in the same frame, the largest motion magnitude in the frame is used to categorize the frame.

	tiny	small	medium	large	huge
UCSD	0	2.85%	26.88%	41.03%	29.24%
Shanghai	0	17.90%	30.34%	29.09%	22.67%
Street	0	75.47%	18.17%	3.17%	3.19%
Live	0	70.70%	24.38%	3.50%	1.43%
UCF	0	66.83%	23.82%	6.55%	2.80%

Table 8: The percentages of anomalous frames in each category under the motion magnitude measurement.

The motion magnitude of anomalies from Street Scene, Live Video, and UCF Crime datatsets tends to be small. As shown in the previous section, the scales of most anomalies in the Street Scene dataset are small, and consequently the motion from these anomalies could also be small. There are large scale anomalies in Live Video dataset and UCF Crime dataset; however, most of the anomalies in both datasets do not introduce large location shift in frames. In contrast, anomalies in UCSD and ShanghaiTech datasets move from one end of the scene to the other, which could give large motion magnitudes.

	OCNN	AE	MIL
p value	0.01	1.3×10^{-5}	0.36
f value	3.792	8.821	1.067

Table 9: ANOVA results for the motion magnitude measurement.

	tiny	small	medium	large	huge
OCNN	0	0.060	0.080	0.077	0.104
AE	0	0.032	0.016	0.075	0.088
MIL	0	0.099	0.102	0.099	0.127

Table 10: The means of AP residuals in each category under the motion magnitude measurement.

Table 9 shows ANOVA results for the small, medium, large, and huge categories. It can be seen that the multiple instance learning method has a p-value 0.36, which indicates that there is no anomaly detection differences among the 4 motion magnitude categories in terms of their group means. This could imply that, as a weakly supervised method, the learning of the MIL is not dominated by the motion magnitude feature. It is also noticeable that different motion magnitudes have impact on the detection performance for the other two unsupervised methods. However, the f value from the one-class neural network is smaller than the autoencoder reconstruction method, which demonstrates that the OCNN method has a better capability to handle the variance of motion magnitude from anomalies. This probably benefits from the OCNN's ability of driving the data representation in the neural network along with its one-class objective.

Table 10 shows the means of the AP residuals in each category. The unsupervised OCNN method has better detection performance than the unsupervised AE method when the motion magnitude is relatively small, while supervised MIL gives the best performance across all 4 categories and has close detection performance among the categories.

4.4 Discussion

During the evaluation when there are multiple anomalies present in the same video frame, the largest value of a measurement is used to determine the category for that frame. This might introduce a bias in the assessment where the true positive frames could be overestimated for the larger anomalies in the datasets. One way to alleviate this problem could be running the evaluation again and using the smallest value to determine the category, then taking both evaluations into consideration in the assessment.

In the assessment we apply dropout to have some variance for each measurement. An alternative way is bootstrapping the datasets for each AP calculation. However, in this alternative way the baseline needs to be computed for every resampling such that the AP residuals can be calculated. Using the auROC metric instead of AP could avoid the baseline calculation for every sample because the area under the baseline of the ROC curve is always 0.5; however, as discussed in Section 1.2, a weak point of auROC is that the metric does not emphasize predictions around the anomaly class.

5 CONCLUSION

In this paper we presented a quartile based quality assessment of video anomaly detection. Quartiles provide information about the spread of the data and statistics based on quartiles are robust to noise in the annotations. Three methods with five datasets were evaluated in terms of the position, scale, and motion magnitude of anomalies. Our evaluation shows anomaly scale introduces detection differences for all three methods evaluated, and both the supervised and unsupervised methods perform better when the anomalies become larger in scale. The assessment also shows the unsupervised autoencoder reconstruction method is immune to the position changes of anomalies while the weakly supervised multiple instance learning method is robust to the motion differences.

The detection performance can also be evaluated with other properties of the anomalies such as the ratio of the anomaly motion magnitude, etc. We hope the framework presented in this paper can inspire other researchers and be used in other computer vision tasks.

ACKNOWLEDGMENTS

This publication has emanated from research supported by Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289_P2, co-funded by the European Regional Development Fund.

REFERENCES

[1] Shuai Bai, Zhiqun He, Yu Lei, Wei Wu, Chengkai Zhu, Ming Sun, and Junjie Yan. 2019. Traffic Anomaly Detection via Perspective Map based on Spatial-temporal Information Matrix. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

[2] Raghavendra Chalapathy, Aditya Krishna Menon, and Sanjay Chawla. 2018. Anomaly Detection using One-Class Neural Networks. *arXiv preprint arXiv:1802.06360* (2018).

[3] Y. Cong, J. Yuan, and J. Liu. 2011. Sparse reconstruction cost for abnormal event detection. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*. 3449–3456. https://doi.org/10.1109/CVPR.2011.5995434

[4] Keval Doshi and Yasin Yilmaz. 2020. Any-Shot Sequential Anomaly Detection in Surveillance Videos. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition CVPR Workshops*. IEEE, 4037–4042. https://doi.org/10.1109/CVPRW50498.2020.00475

[5] Shikha Dubey, Abhijeet Boragule, and Moongu Jeon. 2019. 3D ResNet with Ranking Loss Function for Abnormal Activity Detection in Videos. *2019 International Conference on Control, Automation and Information Sciences (ICCAIS)* (2019), 1–6.

[6] M. Hasan, J. Choi, J. Neumann, A. K. Roy-Chowdhury, and L. S. Davis. 2016. Learning Temporal Regularity in Video Sequences. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 733–742. https://doi.org/10.1109/CVPR.2016.86

[7] Ammar Mansoor Kamoona, Amirali Khodadadian Gostar, Alireza Bab-Hadiashar, and Reza Hoseinnezhad. 2020. Multiple Instance-Based Video Anomaly Detection using Deep Temporal Encoding-Decoding. *CoRR* abs/2007.01548 (2020). arXiv:2007.01548 https://arxiv.org/abs/2007.01548

[8] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. 2014. Large-Scale Video Classification with Convolutional Neural Networks. In *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*. 1725–1732. https://doi.org/10.1109/CVPR.2014.223

[9] Roberto Leyva, Victor Sanchez, and Chang Tsun Li. 2017. The LV dataset: A realistic surveillance video dataset for abnormal event detection. In *Proceedings of the 2017 5th International Workshop on Biometrics and Forensics (IWBF 2017)*. 1–6. https://doi.org/10.1109/IWBF.2017.7935096

[10] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *European Conference on Computer Vision (ECCV)* (2014-01-01). Zürich. /se3/wp-content/uploads/2014/09/coco_eccv.pdf,http://mscoco.org

[11] Kun Liu and Huadong Ma. 2019. Exploring Background-Bias for Anomaly Detection in Surveillance Videos. In *Proceedings of the 27th ACM International Conference on Multimedia*. 1490–1499.

[12] Wen Liu, Weixin Luo, Dongze Lian, and Shenghua Gao. 2018. Future Frame Prediction for Anomaly Detection - A New Baseline. In *2018 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE Computer Society, 6536–6545. https://doi.org/10.1109/CVPR.2018.00684

[13] 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)*. 2720–2727. https://doi.org/10.1109/ICCV.2013.338

[14] Vijay Mahadevan, Wei-Xin LI, Viral Bhalodia, and Nuno Vasconcelos. 2010. Anomaly Detection in Crowded Scenes. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*. 1975–1981.

[15] B. Ramachandra, M. Jones, and R. Vatsavai. 5555. A Survey of Single-Scene Video Anomaly Detection. *IEEE Transactions on Pattern Analysis & Machine Intelligence* 01 (5555), 1–1. https://doi.org/10.1109/TPAMI.2020.3040591

[16] Bharathkumar Ramachandra and Michael J. Jones. 2020. Street Scene: A new dataset and evaluation protocol for video anomaly detection. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*. 2569–2578. https://doi.org/10.1109/WACV45572.2020.9093457

[17] Mahdyar Ravanbakhsh, Moin Nabi, Enver Sangineto, Lucio Marcenaro, Carlo S. Regazzoni, and Nicu Sebe. 2017. Abnormal event detection in videos using generative adversarial nets. In *2017 IEEE International Conference on Image Processing, ICIP*. 1577–1581. https://doi.org/10.1109/ICIP.2017.8296547

[18] Lukas Ruff, Robert A. Vandermeulen, Nico Görnitz, Lucas Deecke, Shoaib A. Siddiqui, Alexander Binder, Emmanuel Müller, and Marius Kloft. 2018. Deep One-Class Classification. In *Proceedings of the 35th International Conference on Machine Learning*, Vol. 80. 4393–4402.

[19] W. Sultani, C. Chen, and M. Shah. 2018. Real-World Anomaly Detection in Surveillance Videos. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6479–6488. https://doi.org/10.1109/CVPR.2018.00678

[20] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning Spatiotemporal Features with 3D Convolutional Networks. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*. 4489–4497. https://doi.org/10.1109/ICCV.2015.510

[21] Hanh Tran and David C. Hogg. 2017. Anomaly Detection using a Convolutional Winner-Take-All Autoencoder. In *BMVC*.

[22] Gaoang Wang, Xinyu Yuan, Aotian Zheng, Hung-Min Hsu, and Jenq-Neng Hwang. 2019. Anomaly Candidate Identification and Starting Time Estimation of Vehicles from Traffic Videos. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 382–390. http://openaccess.thecvf.com/content_CVPRW_2019/html/AI_City/Wang_Anomaly_Candidate_Identification_and_Starting_Time_Estimation_of_Vehicles_from_CVPRW_2019_paper.html