MITSUBISHI ELECTRIC RESEARCH LABORATORIES
http://www.merl.com

Street Scene: A new dataset and evaluation protocol for video anomaly detection

Jones, M.J.; Ramachandra, B.

TR2018-188 January 19, 2019

Abstract

Progress in video anomaly detection research is currently slowed by small datasets that lack a wide variety of activities as well as flawed evaluation criteria. This paper aims to help move this research effort forward by introducing a large and varied new dataset called Street Scene, as well as two new evaluation criteria that provide a better estimate of how an algorithm will perform in practice. In addition to the new dataset and evaluation criteria, we present two variations of a novel baseline video anomaly detection algorithm and show they are much more accurate on Street Scene than two state-of-the-art algorithms from the literature.

arXiv

Street Scene: A new dataset and evaluation protocol for video anomaly detection

Bharathkumar Ramachandra
North Carolina State University
Raleigh, NC
bramach2@ncsu.edu

Michael J. Jones
Mitsubishi Electric Research Labs
Cambridge, MA
mjones@merl.com

Abstract

Progress in video anomaly detection research is currently slowed by small datasets that lack a wide variety of activities as well as flawed evaluation criteria. This paper aims to help move this research effort forward by introducing a large and varied new dataset called Street Scene, as well as two new evaluation criteria that provide a better estimate of how an algorithm will perform in practice. In addition to the new dataset and evaluation criteria, we present two variations of a novel baseline video anomaly detection algorithm and show they are much more accurate on Street Scene than two state-of-the-art algorithms from the literature.

1. Introduction

Surveillance cameras are ubiquitous, and having humans monitor them constantly is not possible. In most cases, almost all of the video from a surveillance camera is unimportant and only unusual video segments are of interest. This is one of the main motivations for developing video anomaly detection algorithms - to automatically find parts of a video that are unusual and flag those for human inspection.

The problem of video anomaly detection is difficult to formulate precisely. One imprecise formulation is as follows. Given one or more training videos from a static camera containing only normal (non-anomalous) events, detect anomalous events in testing video from the same static camera. Providing training video of normal activity is necessary to define what is normal for a particular scene. By *anomalous event*, we mean a spatially and temporally localized segment of video that is significantly different from anything occurring in the training video. What exactly is meant by "significantly different" is difficult to specify and really depends on the target application. This "difference" could be caused by several factors, most commonly unusual appearance or motion of objects in the video. In practice, in the research community, anomalous events in a particular dataset are determined by the dataset creator. Thus, the design and labeling of a dataset for video anomaly detection must be done thoughtfully and carefully.

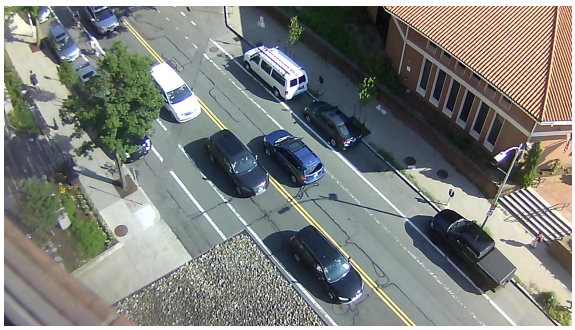

Figure 1. A normal frame from the Street Scene dataset.

After working on this problem, we think there are deficiencies in existing datasets for video anomaly detection. These deficiencies include the simplicity of the scenes for many datasets, the small number of anomalous events, the lack of variety in anomalous events, the very low resolution of some datasets, existence of staged anomalies in some cases, inconsistency in annotation, and the lack of spatial ground truth (in addition to temporal) in some cases. Furthermore, the evaluation criteria that have become standard practice for video anomaly detection have problems. Namely, the evaluation criteria do not properly evaluate spatial localization and do not properly count false positives. In short, they do not give a realistic picture of how an algorithm will perform in practice.

The goal of this paper is to shift the focus of video anomaly detection research to more realistic datasets and more useful evaluation criteria. To this end, we introduce a new dataset for video anomaly detection, called Street Scene, that has more labeled anomalous events and a greater variety of anomalies than previous datasets. Street Scene contains video of a two-way urban street including bike lanes and pedestrian sidewalks (see Figure 1). The video is higher resolution and captures a scene with more varied activity than previous datasets. We also suggest two

new evaluation criteria which we believe give a more accurate picture of how video anomaly detection algorithms will perform in practice than the existing criteria. Finally, we describe two variations of a novel algorithm which greatly outperform two state-of-the-art algorithms on Street Scene.

2. Existing Datasets and Evaluation Criteria

There are a handful of publicly available datasets used to evaluate video anomaly detection algorithms. We discuss each of these below and summarize them in table 1.

2.1. UCSD Pedestrian

The most widely used video anomaly detection dataset is the UCSD pedestrian anomaly dataset [9] which consists of video from two different static cameras (labeled Ped1 and Ped2), each looking at a pedestrian walkway. The Ped1 videos consist of 34 training videos and 36 testing videos each of resolution 238 x 158 pixels. Ped2 consists of 16 training and 12 testing videos of resolution 360 x 240 pixels. Each video contains from 120 to 200 frames. The normal training videos contain groups of people walking along a pedestrian walkway. There are 54 labeled anomalous events (tracks) in Ped1 and 23 in Ped2. The test videos contain 5 different types of anomalies: "bike", "skater", "cart", "walk across", and "other". Anomalous frames are marked for all testing videos, and a subset of the testing videos have pixelwise spatial labels for anomalies per frame.

Despite being widely used, this dataset has various deficiencies. One is that it is modest size, both in terms of number of frames and total anomalies. It only contains 5 different types of anomalies. Another is that all of the anomalies can be detected by only analyzing a single frame at a time. In other words, none of the anomalies really involve any actions evolving over time. Finally, spatial annotations are only provided for some of the testing videos.

2.2. Subway

The Subway dataset [1] contains two long videos of a subway entrance and exit that mainly capture people entering and leaving through turnstiles. Anomalous activities include people jumping or squeezing around the turnstiles, walking the wrong direction, and a person cleaning the walls. Because only two long videos are provided, there are various ambiguities with this dataset such as what frame rate to extract frames, which frames to use as train/test and exactly which frames are labeled as anomalous. Also, there are no spatial ground truth labels. In total, 66 anomalous events are labeled temporally.

2.3. CUHK Avenue

Another widely used dataset is called CUHK Avenue [12]. This dataset consists of short video clips taken from a single outdoor surveillance camera looking at the side of a build with a pedestrian walkway in front of it. The main activity consists of people walking and going into or out of the building. There are 16 training videos and 21 testing videos each of resolution 640 x 360 pixels. The testing videos contain 47 total anomalous events. Anomalies are mostly staged and consist of actions such as a person throwing papers or a backpack into the air, or a child skipping across the walkway. Spatial and temporal anomaly annotations are provided. Like UCSD, this dataset also has a small number and variety of anomalies. In addition since many of the anomalies are staged, they do not seem natural.

2.4. UMN

The UMN dataset contains 11 short clips of 3 scenes of people meandering around an outdoor field, an outdoor courtyard, or an indoor foyer. In each of the clips the anomaly consists of all of the people suddenly running away, hinting at a frantic evacuation scenario. The scene is staged and there is one anomalous event per clip. There is no clear specification of a split between training and testing frames and anomalies are only labeled temporally.

2.5. ShanghaiTech Campus

A recent paper by Liu et al. [11] introduced a new dataset for video anomaly detection called ShanghaiTech Campus. It consists of 13 different training scenes (12 of which are used for testing) and 317,398 total frames (274,515 for training and 42,883 for testing). Each color frame has resolution 856 x 480 pixels. A typical video shows people walking along a sidewalk. Anomalies include bikers, skateboarders, people running or chasing, and people fighting. There are 130 anomalous events in total which are labeled both spatially and temporally.

This dataset is a good addition to the field, but it still has only a modest number of anomalous events (130) and papers that have evaluated on it [13, 11] still use the frame-level and pixel-level criteria introduced in [9] for which there are problems that we discuss below. Furthermore, papers that have used this dataset have trained a single model on all 13 different training scenes. In our view this does not fit with the formulation of video anomaly detection because an event that is anomalous in one scene (such as a person running) may not be anomalous in a second scene since training videos in the second scene describing normality may include running people whereas the first does not. Thus, different models are necessary for each different scene. This is not a deficiency in the dataset itself, but does set a precedent on how to use the dataset.

2.6. Evaluation Criteria

Almost every recent paper for video anomaly detection [15, 16, 22, 7, 19, 17, 4, 14, 23, 21, 24, 6, 5, 20, 12, 18, 10,

Dataset	Total Frames	Training Frames	Avg Frames per Training Video	Testing Frames	Avg Frames per Testing Video	Anomalous Events
UCSD Ped1 and Ped2*	18,560	9,350	187	9,210	192	77
Subway	139 min	25 min	N/A	114 min	N/A	66
CUHK Avenue	30,652	15,328	958	15,324	730	47
UMN**	4 min 17 sec	N/A	N/A	N/A	N/A	11
ShanghaiTech***	317,398	274,515	832	42,883	401	130
Street Scene	**203,257**	**56,847**	**1,235**	**146,410**	**4,183**	**203**

Table 1. Characteristics of video anomaly detection datasets. *aggregates from 2 cameras. **aggregates from 3 cameras. ***aggregates from 13 cameras.

2, 3] has used one or both of the evaluation criteria specified in Li et al. [9] which also introduced the UCSD pedestrian dataset. The first criterion, referred to as the *frame-level* criterion, counts a frame with any detected anomalous pixels as a positive frame and all other frames as negative. The frame-level ground truth annotations are then used to determine which detected frames are true positives and which are false positives, thus yielding frame-level true positive and false positive rates. This criterion uses no spatial localization and counts a frame as a correct detection (true positive) even if the detected anomalous pixels do not overlap with any ground truth anomalous pixels. Even the authors who proposed this criterion stated that they did not think it was the best one to use [9]. We have observed that some methods that claim state-of-the-art performance on frame-level criterion perform poor spatial localization in practice.

The other criterion is the *pixel-level* criterion and tries to take into account the spatial locations of anomalies. Unfortunately, it does so in a problematic way. The pixel-level criterion still counts true positive and false positive frames as opposed to true and false positive anomalous regions. A frame with ground truth anomalies is counted as a true positive detection if at least 40% of its ground truth anomalous pixels are detected. Any frame with no ground truth anomalies is counted as a false positive frame if at least one pixel is detected as anomalous. This criterion has serious deficiencies. For example, anytime an algorithm detects a single pixel of a frame as anomalous, it might as well label all pixels of that frame as anomalous. This would guarantee a correct detection if the frame has a ground truth anomaly and would not further increase the false positive rate if it does not. That is, it does not reward tightness of localization or penalize looseness of it. Furthermore, it does not give a realistic measure of how many false positive regions to expect an algorithm to have in practice. This is because false positive regions are not even counted for frames containing ground truth anomalies, and a frame with no ground truth anomalies can only have a single false positive even if an algorithm falsely detects many different false positive regions in that frame.

Better evaluation criteria are clearly needed.

3. Description of Street Scene

The Street Scene dataset consists of 46 training video sequences and 35 testing video sequences taken from a static USB camera looking down on a scene of a two-lane street with bike lanes and pedestrian sidewalks. See Figure 1 for a typical frame from the dataset. Videos were collected from the camera at various times during two consecutive summers. All of the videos were taken during the daytime. The dataset is challenging because of the variety of activity taking place such as cars driving, turning, stopping and parking; pedestrians walking, jogging and pushing strollers; and bikers riding in bike lanes. In addition the videos contain changing shadows, and moving background such as a flag and trees blowing in the wind.

There are a total of 203,257 color video frames (56,847 for training and 146,410 for testing) each of size 1280 x 720 pixels. The frames were extracted from the original videos at 15 frames per second.

We wanted the dataset to contain only "natural" anomalies, i.e. not staged by "actors". Since anomalies are determined by what is not in the training video, we tried to be thoughtful about what to include in training. To this end, the training sequences were chosen to meet the following criteria:

- If people are present, they are walking, jogging or pushing a stroller in one direction on a sidewalk; or they are getting into or out of their car including walking along the side of their car; or they are stopped in front of a parking meter.
- If a car is present, it is legally parked; or it is driving in the appropriate direction in a car lane; or stopped in a car lane due to traffic; or making a legal turn across traffic; or leaving/entering a parking spot on the side of the street.
- If bikers are present, they are riding in the appropriate direction in a bike lane; or turning from the intersecting road into a bike lane or from a bike lane onto the intersecting road.

These criteria for normal activity imply that the following activities, for example, are anomalous and thus do not appear in the training videos: Pedestrians walking across the

Anomaly Class	Instances	Anomaly Class	Instances
1. Jaywalking	60	10. Car illegally parked	5
2. Biker outside lane	42	11. Person opening trunk	4
3. Loitering	37	12. Person exits car on street	3
4. Dog on sidewalk	11	13. Skateboarder in bike lane	2
5. Car outside lane	9	14. Person sitting on bench	2
6. Worker in bushes	8	15. Metermaid ticketing car	1
7. Biker on sidewalk	7	16. Car turning from parking space	1
8. Pedestrian reverses direction	5	17. Motorcycle drives onto sidewalk	1
9. Car u-turn	5		

Table 2. Anomaly classes and number of instances of each in the Street Scene dataset.

road (i.e. jaywalking), pedestrians stopped on the sidewalk (loitering), pedestrians walking one direction and then turning around and walking the opposite direction, bikers on the sidewalk, bikers outside a bike lane (except when turning into a bike lane from the intersecting street) cars making u-turns, cars parked illegally, cars outside a car lane (except when turning or parked, parking or leaving a parking spot).

The 35 testing sequences have a total of 203 anomalous events consisting of 17 different anomaly types. A complete list of anomaly types and the number of each in the test set is given in Table 2.

Ground truth annotations are provided for each testing video in the form of bounding boxes around each anomalous event in each frame. Each bounding box is also labeled with a track number, meaning each anomalous event is labeled as a track of bounding boxes. Track lengths vary from tens of frames to 5200 which is the length of the longest testing sequence. A single frame can have more than one anomaly labeled.

Labeling anomalies is inherently ambiguous. When exactly does an anomaly such as jaywalking or a car making a u-turn begin and end? How far outside the bike lane does a biker need to be to constitute a "biker outside lane" anomaly? If two pedestrians are holding hands while walking, is that normal even though this didn't occur in any training sequences? What if this occurred in training on one sidewalk but not the other? The list could go on.

In short, we tried to use common sense when such issues come up during labeling. We decided to start labeling jaywalking on the frame where the person leaves the curb and goes into the street. A biker needs to be all the way outside the bike lane (not touching the lane line) to be counted as anomalous. Pedestrians holding hands are not different enough from pedestrians walking side by side to be counted as anomalous (especially at the low resolution of pedestrians in StreetScene). These inherent ambiguities also inform our evaluation criteria which are described next.

The Street Scene dataset is available for download on MERL's website: www.merl.com.

4. New Evaluation Criteria

As discussed in Section 2.6, the main criteria used by previous work to evaluate video anomaly detection accuracy have significant problems. A good evaluation criterion should provide a good idea of the fraction of anomalies an algorithm can detect and the number of false positive regions an algorithm can be expected to mistakenly find per frame. To this end, we propose two new criteria for evaluating algorithms on Street Scene which are inspired by evaluation criteria that are commonly used for object detection.

Our new evaluation criteria are informed by the following considerations. Similar to object detection criteria, using the intersection over union (IOU) between a ground truth anomalous region and a detected anomalous region for determining whether an anomaly is detected is a good way to insure rough spatial localization. For video anomaly detection, the IOU threshold should be low to allow some imprecision in localization because of issues like imprecise labeling (bounding boxes) and the fact that some algorithms detect anomalies that are close to each other as one large anomalous region which shouldn't be penalized. Similarly, shadows may cause larger anomalous regions than what are labeled (in Street Scene only anomalous objects are included in the ground truth anomalous bounding box, not the object's shadow). This is another gray area in labeling. One could reasonably argue that the shadow of a jaywalker, for example, should be included in the ground truth bounding box. We don't think such larger than expected anomalous-region detections should be penalized. We use an IOU threshold of 0.1 in our experiments.

Also, because a single frame can have multiple ground-truth anomalous regions, correct detections should be counted at the level of an anomalous region and not at the level of a frame.

False positives should be counted for each falsely detected anomalous region, i.e. by each detected anomalous region that does not significantly overlap with a ground truth anomalous region. This allows more than one false positive per frame and also false positives in frames with ground truth annotations, unlike the previous criteria.

In practice, for an anomaly that occurs over many frames, it is important to detect the anomalous region in at least some of the frames, but it is usually not important to detect the region in every frame in the track. This is especially true considering the ambiguities for when to begin and end an anomalous track mentioned earlier. A good anomaly detection algortithm should detect every anomalous event (which occurs over many frames) but it can do this by detecting the anomalous event region in only some of the frames. Because the Street Scene dataset provides track numbers for each anomalous region which uniquely identify the event to which an anomalous region belongs, it is easy to compute such a criterion.

4.1. Track-Based Detection Criterion

The track-based detection criterion measures the track-based detection rate (TBDR) versus the number of false positive regions per frame.

A ground truth track is considered detected if at least a fraction α of the ground truth regions in the track are detected.

A ground truth region in a frame is considered detected if the intersection over union (IOU) between the ground truth region and a detected region is greater than or equal to β.

$$\text{TBDR} = \frac{\text{num. of anomalous tracks detected}}{\text{total num. of anomalous tracks}}. \quad (1)$$

A detected region in a frame is a false positive if the IOU between it and every ground truth region in that frame is less than β.

$$\text{FPR} = \frac{\text{total false positive regions}}{\text{total frames}} \quad (2)$$

where FPR is the false-positive rate per frame.

Note that a single detected region can cover two or more different ground truth regions so that each ground truth region is detected (although this is rare).

In our experiments below, we use $\alpha = 0.1$ and $\beta = 0.1$.

4.2. Region-Based Detection Criterion

The region-based detection criterion measures the region-based detection rate (RBDR) over all frames in the test set versus the number of false positive regions per frame.

As with the track-based detection criterion, a ground truth region in a frame is considered detected if the intersection over union (IOU) between the ground truth region and a detected region is greater than or equal to β.

$$\text{RBDR} = \frac{\text{num. of anomalous regions detected}}{\text{total num. of anomalous regions}}. \quad (3)$$

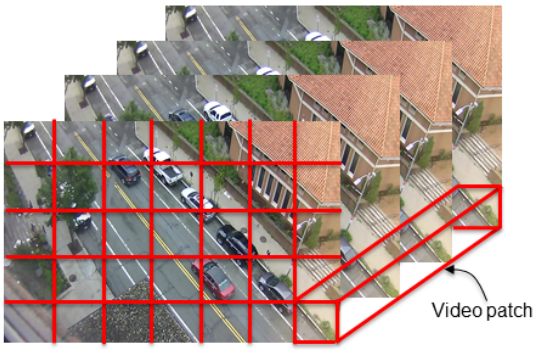

Figure 2. Illustration of a grid of regions partitioning a video frame and a video patch encompassing 4 frames. This figure show non-overlapping regions, but in our experiments we use overlapping regions.

The RBDR is computed over all anomalous regions in all frames of the test set.

The number of false positives per frame is calculated in the same way as with the track-based detection criterion.

As with any detection criterion, there is a trade-off between detection rate (true positive rate) and false positive rate which can be captured in a ROC curve computed by changing the threshold on the anomaly score that determines which regions are detected as anomalous.

When a single number is desired, we suggest summarizing the performance with the average detection rate for false positive rates from 0 to 1, i.e. the area under the ROC curve for false positive rates less than or equal to 1.

5. Baseline Algorithms

We describe two variations of a novel algorithm for video anomaly detection which we evaluate along with two previously published algorithms on the Street Scene dataset in Section 6. The new algorithm is very straightforward and is based on dividing the video into spatio-temporal regions which we call video patches, storing a set of exemplars to represent the variety of video patches occuring in each region, and then using the distance from a testing video patch to the nearest neighbor exemplar as the anomaly score.

First, each video is divided into a grid of spatio-temporal regions of size $H \times W \times T$ pixels with spatial step size s and temporal step size 1 frame. In the experiments in Section 6 we choose H=40 pixels, W=40 pixels, T=4 or 7 frames, and s = 20 pixels. See Figure 2 for an illustration.

The baseline algorithm has two phases: a training or model-building phase and a testing or anomaly detection phase. In the model-building phase, the training (normal) videos are used to find a set of video patches (represented by feature vectors described later) for each spatial region that represent the variety of activity in that spatial region. We call these representative video patches, exemplars. In

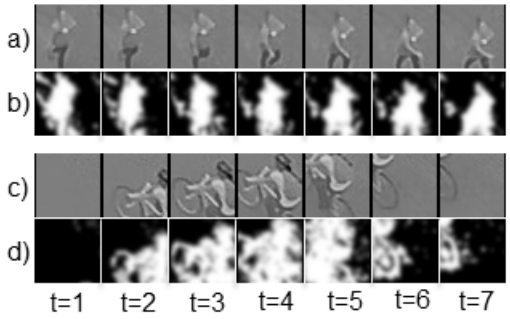

Figure 3. Example blurred FG masks which are concatenated and vectorized into a feature vector. a and c show two video patches consisting of 7 frames cropped around a spatial region. b and d show the corresponding blurred FG masks.

the anomaly detection phase, the testing video is split into the same regions used during training and for each testing video patch, the nearest exemplar from its spatial region is found. The distance to the nearest exemplar serves as the anomaly score.

The only differences between the two variations are the feature vector used to represent each video patch and the distance function used to compare two feature vectors.

The foreground (FG) mask variation uses blurred FG masks for each frame in a video patch. The FG masks are computed using a background (BG) model that is updated as the video is processed. The BG model used in the experiments is a very simple mean color value per pixel. The BG model is initialized with the mean of the first 200 frames of the input video and is then updated as follows:

$$B_{t+1} = (19 * B_t + I_t)/20 \tag{4}$$

where B_t is the current BG model at time t and I_t is the input video frame at time t. The constants in the equation were set empirically.

Given the BG image, B_t, and input frame, I_t at time t, the FG mask at time t is computed as

$$FG_t(i,j) = \begin{cases} 1 & \text{if } |BG_t^C(i,j) - I_t^C(i,j)| > \Theta \\ & \text{for all channels in } C \\ 0 & otherwise \end{cases} \tag{5}$$

where $C \in \{R, G, B\}$ for color images and $C \in \{gray\}$ for gray-scale images. Indices i and j are the row and column indices to a pixel. The threshold, Θ, was found empirically. In the experiments, $\Theta = 12$ (for color values from 0 to 255).

The FG mask is then blurred using a Gaussian kernel to make the L_2 distance between FG masks more robust. The FG mask feature vector is formed by concatenating all of the blurred FG masks from all frames in a video patch and then vectorizing (see Figure 3).

The flow-based variation uses optical flow fields computed between consecutive frames in place of FG masks. The flow fields within the region of each video patch frame are concatenated and then vectorized to yield a feature vector twice the length of the feature vector from the FG mask baseline (due to the dx and dy components of the flow field). In our experiments we use the optical flow algorithm of Kroeger et al. [8] to compute flow fields.

In the model building phase, a distinct set of exemplars is selected to represent normal activity in each spatial region. Our exemplar selection method is straightforward. For a particular spatial region, the exemplar set is initialized to the empty set. We slide a spatial-temporal window (with step size equal to one frame) along the temporal dimension of each training video to give a series of video patches which we represent by either a FG-mask based feature vector or a flow-based feature vector depending on the algorithm variation as described above. For each video patch, we compare it to the current set of exemplars. If the distance to the nearest exemplar is less than a threshold then we discard that video patch. Otherwise we add it to the set of exemplars.

The distance function used to compare two exemplars depends on the feature vector. For blurred FG mask feature vectors, we use L_2 distance. For flow-field feature vectors we use normalized L_1 distance:

$$dist(\mathbf{u}, \mathbf{v}) = \sum_i \frac{|u_i - v_i|}{|u_i| + |v_i| + \epsilon} \tag{6}$$

where u and v are two flow-based feature vectors and ϵ is a small positive constant used to avoid division by zero.

Given a model of normal video which consists of a different set of exemplars for each spatial region of the video, the anomaly detection is simply a series of nearest neighbor lookups. For each spatial region in a sequence of T frames of a testing video, compute the feature vector representing the video patch and then find the nearest neighbor in that region's exemplar set. The distance to the closest exemplar is the anomaly score for that video patch.

This yields an anomaly score per overlapping video patch. These are used to create a per-pixel anomaly score matrix for each frame. The anomaly score for a video patch is stored in the middle frame for that set of T frames. The first $T/2-1$ frames and the last $T/2+1$ frames of the testing video are not assigned any anomaly scores from video patches and thus get all 0's. A pixel covered by two or more video patches is assigned the average score from all video patches that include the pixel.

When computing ROC curves according to either of the track-based or region-based criteria, for a given threshold, all *pixels* with anomaly scores above the threshold are labeled anomalous. Then anomalous *regions* are found by computing the connected components of anomalous pixels. These anomalous regions are compared to the ground truth regions according to one of the above criteria.

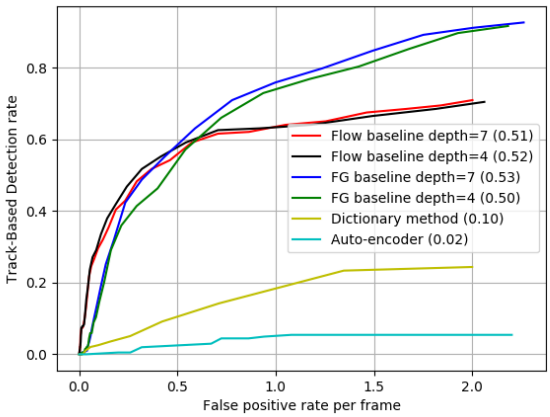

Figure 4. ROC curves for track-based criterion for different methods.

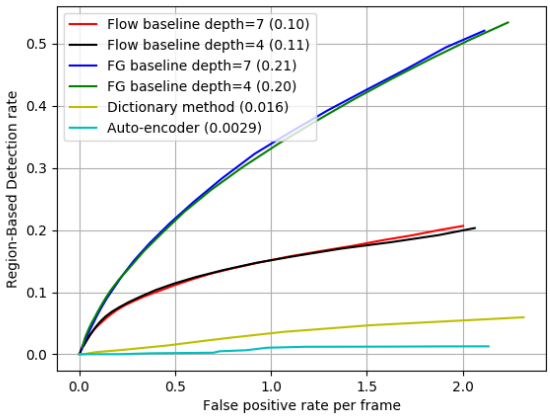

Figure 5. ROC curves for region-based criterion for different methods.

6. Experiments

In addition to the two variations of our baseline video anomaly detection method, we also tested two previously published methods that do very well on other publicly available datasets. The first is the dictionary method of Lu et al. [12] which fits a sparse combination of dictionary basis feature vectors to a feature vector representing each spatio-temporal window of the test video. A dictionary of basis feature vectors is learned from the normal training videos for each spatial region independently. This method reported state-of-the-art results on UCSD, Subway and CUHK Avenue datasets. Code was provided by the authors.

The second method is from Hasan et al. [6] which uses a deep network auto-encoder to learn a model of normal frames. The anomaly score for each pixel is the reconstruction error incurred by passing a clip containing this pixel through the auto-encoder. The assumption is that anomalous regions of a frame will not be reconstucted well by the auto-encoder. This method is also competitive with other state-of-the-art results on standard datasets and evaluation criteria. We used our own implementation of this method.

Figures 4 and 5 show ROC curves for our baseline methods as well as the dictionary and auto-encoder methods on Street Scene using the newly proposed track-based and region-based criteria. The numbers in parentheses for each method in the figure legends are the areas under the curve for false positive rates from 0 to 1. Clearly, the dictionary and auto-encoder methods perform poorly on Street Scene. Our baseline methods do much better although there is still much room for improvement.

The dictionary method achieves computational efficiency by limiting the dictionary and by learning sparse sets of dictionary members that may be combined together from the training videos. This idea seems to work well on other, smaller datasets, but on the larger and more varied Street Scene it seems to restrict the expressiveness of the model too much so that many normal testing video patches are not well reconstructed by the model.

The auto-encoder method gives per-pixel anomaly scores (not per-region). This results in many isolated anomalous pixels or small clusters of pixels which in turn results in an explosion of anomalous regions when connected components of anomalous pixels are computed. A set of anomalous regions per frame is needed to compute the track-based and region-based criteria. To alleviate this problem, we post-process the pixelwise anomaly scores using a sliding window box filter. This filter checks if the number of anomalous pixels (given an anomaly score threshold) within it is above 30% of the total number of pixels in the filter and if so all pixels within the box filter are labeled "anomalous". Otherwise all pixels are labeled "normal". This removes small clusters of anomalous pixels and merges together larger clusters of anomalous pixels that may not actually be connected. This greatly improves the ROC curves for the auto-encoder method. In the results shown a 40x40 box filter is used with a step size of 20. Even with this post-processing, the auto-encoder results are poor on Street Scene. This is most likely due to the huge variety of normal variations that are present in the training videos that the auto-encoder is not able to model well.

Our baseline algorithms perform reasonably well on Street Scene. They store a large set of exemplars (typically between 1000 and 3000 exemplars) in regions where there is a lot of activity such as the street, sidewalk and bike lane regions. On other regions such as the building walls or roof tops, only a single exemplar is stored.

For the two baseline variations using the track-based criteria, the flow-based method does best for low false-positive rates (arguably the most important part of the ROC curve).

The flow field provides more useful information than FG masks for most of the anomalies (the main exception being loitering anomalies which are discussed below). The FG-based method does better using the region-based criterion. The number of frames used in a video patch (4 or 7) does not have a large effect on either variation.

The baseline algorithms do best at detecting anomalous activities such as jaywalking, illegal u-turn, and bikers or cars outside their lanes because these anomalies have distinctive motions compared to the typical motions in the regions where they occur.

The loitering anomalies (and other largely static anomalies such as illegally parked cars) are the most difficult for the baseline methods because they do not contain any motion except at the beginning in which a walking person transitions to loitering. For the flow-based method, the loitering anomalies are completely invisible. For the FG-based method, the beginning of the loitering anomaly is visible since the BG model takes a few frames to absorb the motionless person. This is the main reason why the flow-based method is worse than the FG-based method for higher detection rates. The FG-based method can detect some of the loitering anomalies while the flow-based method cannot.

A similar effect explains the region-based results in which the FG-based method does better than the flow-based method. The loitering and other "static" anomalies make up a disproportionate fraction of the total anomalous regions because many of them occur over many frames. The FG-based method detects some of these regions while the flow-based method misses essentially all of them. So even though the flow-based method detects a greater fraction of all anomalous *tracks* (at low false positive rates) it detects a smaller fraction of all anomalous *regions*. This also suggests that the track-based criterion may provide a better measure of how an algorithm will perform in practice.

Some visualizations of the detection results for the flow-based method (using T=4) are shown in Figures 6 - 8. In the figures, red tinted pixels are anomaly detections and blue boxes show the ground truth annotations. Figure 6 shows the correct detection of a motorcycle that rides onto a sidewalk. Figure 7 shows the correct detection of a jaywalker. Note that the anomalous region detected around the jaywalker includes the person's shadow which is arguably also anomalous. The ground truth bounding box does not include the shadow, but this is still counted as a correct detection because of the low IOU threshold (0.1). There is also a false positive region near the bottom, left corner which is due to a person's shadow. Figure 8 shows the correct detection of a car making an illegal u-turn. These correct detections show the range of scales that the baseline algorithm can handle despite only looking at a single scale of 40x40 pixel video patches.

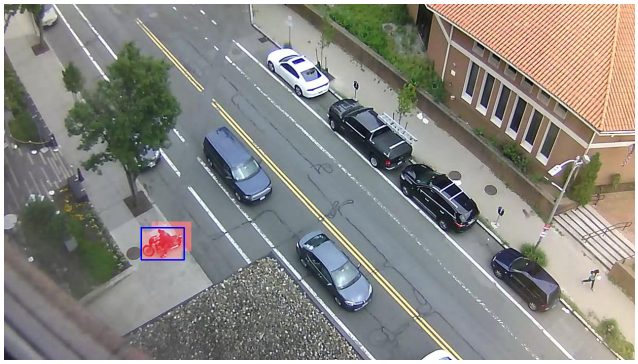

Figure 6. Detection result for Flow baseline showing correctly detected motorcycle driving onto the sidewalk.

Figure 7. Detection result for Flow baseline showing correctly detected jaywalker as well as a false positive on a person's shadow.

Figure 8. Detection result for Flow baseline showing correctly detected illegal u-turn.

7. Conclusions

We have presented a new dataset and new evaluation criteria for video anomaly detection that we hope will help to spur new innovations in this field. The Street Scene dataset has more anomalous events and is a more complex scene than currently available datasets. It will be made publicly available. The new evaluation criteria fix the problems with the criteria typically used in this field, and will give a more realistic idea of how well an algorithm performs in practice.

In addition, we have presented two variations of a new video anomaly detection algorithm that is straightfor-

ward and outperforms two previously published algorithms which do well on previous datasets but not on Street Scene. The new nearest-neighbor based algorithms may form an interesting foundation to build on.

References

[1] A. Adam, E. Rivlin, I. Shimshoni, and D. Reinitz. Robust real-time unusual event detection using multiple fixed-location monitors. *PAMI*, 2008. 2

[2] B. Antic and B. Ommer. Video parsing for abnormality detection. pages 2415–2422. IEEE, Nov. 2011. 3

[3] B. Anti and B. Ommer. Spatio-temporal Video Parsing for Abnormality Detection. *arXiv preprint arXiv:1502.06235*, 2015. 3

[4] K.-W. Cheng, Y.-T. Chen, and W.-H. Fang. Video anomaly detection and localization using hierarchical feature representation and gaussian process regression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2909–2917, 2015. 3

[5] Y. Cong, J. Yuan, and J. Liu. Abnormal event detection in crowded scenes using sparse representation. *Pattern Recognition*, 46(7):1851–1864, July 2013. 3

[6] M. Hasan, J. Choi, J. Neumann, A. Roy-Chowdhury, and L. Davis. Learning temporal regularity in video sequences. In *CVPR*, 2016. 3, 7

[7] L. Kratz and K. Nishino. Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1446–1453. IEEE, 2009. 3

[8] T. Kroeger, R. Timofte, D. Dai, and L. V. Gool. Fast optical flow using dense inverse search. In *ECCV*, 2016. 6

[9] W. Li, V. Mahadevan, and N. Vasconcelos. Anomaly detection and localization in crowded scenes. *PAMI*, 2014. 2, 3

[10] W. Liu, W. Luo, D. Lian, and S. Gao. Future Frame Prediction for Anomaly Detection – A New Baseline. *arXiv:1712.09867 [cs]*, Dec. 2017. arXiv: 1712.09867. 3

[11] W. Liu, W. Luo, D. Lian, and S. Gao. Future frame prediction for anomaly detection - a new baseline. In *CVPR*, 2018. 2

[12] C. Lu, J. Shi, and J. Jia. Abnormal event detection at 150 fps in matlab. In *ICCV*, 2013. 2, 3, 7

[13] W. Luo, W. Liu, and S. Gao. A revisit of sparse coding based anomaly detection in stacked rnn framework. In *ICCV*, 2017. 2

[14] K. Ma, M. Doescher, and C. Bodden. Anomaly Detection In Crowded Scenes Using Dense Trajectories. 3

[15] V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos. Anomaly detection in crowded scenes. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1975–1981, June 2010. 3

[16] R. Mehran, A. Oyama, and M. Shah. Abnormal crowd behavior detection using social force model. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 935–942. IEEE, 2009. 3

[17] M. Sabokrou, M. Fathy, M. Hoseini, and R. Klette. Real-time anomaly detection and localization in crowded scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 56–62, 2015. 3

[18] M. Sabokrou, M. Fayyaz, M. Fathy, and R. Klette. Deep-Cascade: Cascading 3d Deep Neural Networks for Fast Anomaly Detection and Localization in Crowded Scenes. *IEEE Transactions on Image Processing*, 26(4):1992–2004, Apr. 2017. 3

[19] M. Sabokrou, M. Fayyaz, M. Fathy, and others. Fully Convolutional Neural Network for Fast Anomaly Detection in Crowded Scenes. *arXiv preprint arXiv:1609.00866*, 2016. 3

[20] V. Saligrama and Z. Chen. Video anomaly detection based on local statistical aggregates. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2112–2119. IEEE, 2012. 3

[21] H. Vu, D. Phung, T. D. Nguyen, A. Trevors, and S. Venkatesh. Energy-based Models for Video Anomaly Detection. *arXiv preprint arXiv:1708.05211*, 2017. 3

[22] Weixin Li, V. Mahadevan, and N. Vasconcelos. Anomaly Detection and Localization in Crowded Scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(1):18–32, Jan. 2014. 3

[23] S. Wu, B. E. Moore, and M. Shah. Chaotic invariants of lagrangian particle trajectories for anomaly detection in crowded scenes. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2054–2060. IEEE, 2010. 3

[24] D. Xu, E. Ricci, Y. Yan, J. Song, and N. Sebe. Learning deep representations of appearance and motion for anomalous event detection. *arXiv preprint arXiv:1510.01553*, 2015. 3