Classifier Two-Sample Test for Video Anomaly Detections

Yusha Liu*
yushal@cs.cmu.edu
Chun-Liang Li*
chunlial@cs.cmu.edu
Barnabás Póczos
bapoczos@cs.cmu.edu

Machine Learning Department
Carnegie Mellon University
Pittsburgh, USA
(* denote equal contribution)

Abstract

In this paper, we study challenging anomaly detections in streaming videos under fully unsupervised settings. Unsupervised unmasking methods [12] have recently been applied to anomaly detection; however, the theoretical understanding of it is still limited. Aiming to understand and improve this method, we propose a novel perspective to establish the connection between the heuristic unmasking procedure and multiple classifier two sample tests (MC2ST) in statistical machine leaning. Based on our analysis of the testing power of MC2ST, we present a history sampling method to increase the testing power as well as to improve the performance on video anomaly detection. We also offer a new frame-level motion feature that has better representation and generalization ability, and obtain improvement on several video benchmark datasets. The code could be found at https://github.com/MYusha/Video-Anomaly-Detection.

1 Introduction

Anomaly detection in video streams is a challenging task, because the definition of anomaly is never perfectly clear and is highly influenced by the context. Therefore, many supervised learning based approaches [3, 6, 14, 19, 21, 25] require large amount of labeled information. Recently, unsupervised learning based methods [7, 12] have also been studied. Without assuming any labeled information is available in advanced, unsupervised learning is more challenging than its supervised learning counterpoint.

There are two major approaches for unsupervised learning anomaly detection, including offline [7] and online [12] algorithms. Del Giorno et al. [7] is related to permutation tests with theoretical statistical guarantees [31]; however, it is usually more time consuming and is not suitable for real-time detection. Ionescu et al. [12] proposed an unmasking method on sliding windows to compare two sets of video frames with the goal of detecting changes. This approach has promising empirical performance, but the underlying theory is not well understood yet.

A key step of Ionescu et al. [12] is to determine whether two given consecutive windows of frames are "similar" or not. If we assume the first window follows distribution $\mathbb{P}$ and the second window follows another distribution $\mathbb{Q}$, then the task is reduced to the two-sample test problem [4, 11, 15, 20, 24, 30, 32]. A two sample hypothesis test aims to compare two distributions via their samples. In this paper, our goal is to understand the connection between two-sample tests and the "unmasking method" for streaming video anomaly detections, provide theoretical justification, and improve empirical performance. Our contributions are four-fold:

- We study the connection between the intuitive *unmasking method* [12] and the *multiple classifier two-sample test* (MC2ST) in statistical machine learning, and show that [12] is a special case of M2CST.
- We study the theory of MC2ST by deriving its asymptotic testing power.
- Based on the derived testing power, we propose a "history sampling" scheme to boost the performance. We demonstrate the improved accuracy of our method on several benchmark datasets.
- Motivated by our theoretical results on the testing power of M2CST, we propose a better motion feature as a frame-level descriptor for videos.

2 Related Works

Video anomaly detections are usually studied under the supervised settings [2, 3, 5, 6, 14, 19, 21, 25]. The most common approach is to model normal activity patterns and detect outliers under pre-defined metrics and flag them as anomaly. There are various approaches for building these models from training data, such as the Social Force Model [25], Gaussian process [3] and probabilistic Latent Semantic Analysis [19]. Building a dictionary from normal events and then detecting anomalies based on the reconstruction cost from the dictionary is also popular [5, 6, 21]. Deep learning approaches with different network architectures are emerging recently, including sparse auto-encoder [27], stacked Recurrent Neural Network [22] and spatio-temporal adversarial networks [17].

Many works rely on the estimation of the normal distribution of events. Assuming that the abnormal events occur less often, Dutta and Banerjee [8], Zhao et al. [35] build the normal model and gradually update it with an unsupervised method. Ren et al. [26] instead use a dictionary of atoms to represent different types of normal behaviors. Kim and Grauman [14] propose to learn a generative model for local activities and use a Markov Random Field graph to estimate normality. Xu et al. [33] employ deep learning features and use several SVMs to detect anomalies based on the learned patterns.

We instead study a completely unsupervised method that requires no training data and normal patterns of any form. Del Giorno et al. [7] aim to detect abnormal events that are independent of temporal order by permuting the frames first and then detect distinguishable parts as anomaly. The work most related to us isIonescu et al. [12], where they propose to process video in an on-line fashion and detect sudden abnormal behaviors using the unmasking method discussed in Section 4.1.

3 Video Anomaly Detections and Two-Sample Test

Different from existing supervised learning works [21, 23, 25, 33], we consider an unsupervised learning setting in video anomaly detection problems, where there is no training

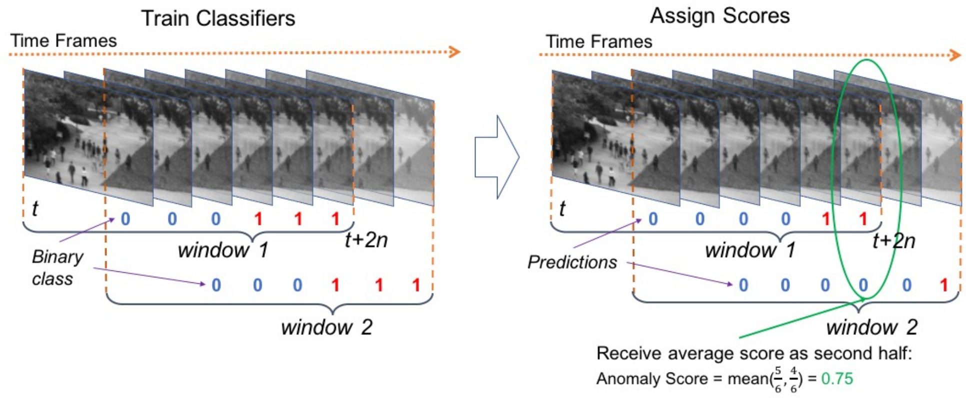


Figure 1: An illustration of the sliding window and average scores.

data with labeling [7, 12]. There are two main frameworks for unsupervised learning detection, including offline [7] and streaming (online) [12]. The offline algorithm [7] requires several permutations of frames to break the time ordering, which is not suitable for real-time applications. In this paper, we consider the online unsupervised learning framework.

Considering that abnormal events often have direct and obvious differences from preceding events, we can use sliding window to detect anomaly in videos. For a sliding window, assuming the first part of the sliding window is normal, we decide that the latter part contains abnormal behaviors if there is a significant difference between them. This framework is also adopted by Ionescu et al. [12], who obtain promising results as an unsupervised learning method. Suppose that we are currently at the window from frame t to $t+2n$, where n denotes the size of each half. The first part of frames $t \sim t+n$ is labeled as class 0, and the second part of frames $t+n+1 \sim t+2n$ as class 1.

We train classifiers on this window and obtain the training accuracy. We then propogate the accuracy to each frame in the second half as the anomaly score, whisch can easily be interpreted as the indicator of how different the two sets are [7, 12]. The sliding window then moves with a stride, which may cause the covering range of consecutive windows to overlap, and we repeat the aforementioned classification procedure again. The final anomaly score of each frame is assigned as the average of all scores it has received as the second half of different sliding windows. An illustration of the sliding window scheme can be found in Figure 1.

3.1 Classifier Two-Sample Test (C2ST)

Without loss of generality, we assume the normal events is followed by a distribution $\mathbb{P}$, while the anomaly is followed by $\mathbb{Q}$. We then reformulate the streaming video anomaly detection decribed above as a *two-sample test* task. We denote frames from two windows $t \sim t+n$ as $\{x_i\}_{i=1}^n$ and frames $t+n+1 \sim t+2n$ as $\{y_i\}_{i=1}^n$, where $\{x_i\}_{i=1}^n \sim \mathbb{P}$ and $\{y_i\}_{i=1}^n \sim \mathbb{Q}$. The null hypothesis is that the two sets of frames are from the same distribution (*e.g.* $\mathbb{P} = \mathbb{Q}$), while the alternative hypothesis is $\mathbb{P} \neq \mathbb{Q}$.

Several two-sample test algorithms have been proposed in statistics and machine learning [4, 11, 15, 20, 24, 30, 32]. Recently, classifier two-sample test is demonstrated to enjoy the advantage in testing high dimensional distributions [10, 20]. Similar ideas have also been applied to video anomaly detection [7, 12]. However, the connection between video anomaly detection and (classifier) two-sample test has not been explicitly established. In this section, we revisit the classifier two-sample test (C2ST) with theoretical justification for Ionescu et al. [12], and a direction to improve these existing works (Section 3.3).

Classifier two-sample test (C2ST) conducts the hypothesis test via training a binary classifier to distinguish two sets of samples, and uses the classification accuracy as the proxy to make the decision. If the accuracy is high, then we are confident to reject $H_0 : \mathbb{P} = \mathbb{Q}$. When we bound the Type-I error[1] with significance level α, we want to maximize the *power* of the test, which is defined as $1-\beta$, where β is the Type-II error[2]. Intuitively speaking, given a tolerance level of being false positive (treat the frame as anomaly when it is not), we want to detect the true anomalies as many as possible. Improving testing power for other tests have been studied [11, 18, 34], however; improving the power of classifier two sample test is not well understood. Next, we show how to improve testing power of C2ST with the connection with Ionescu et al. [12].

3.2 Increasing Testing Power via Multiple Classifiers

If $\{x_i\}_{i=1}^n \sim \mathbb{P}$ and $\{y_i\}_{i=1}^n \sim \mathbb{Q}$, the goal is to test the null hypothesis $H_0 : \mathbb{P} = \mathbb{Q}$. We construct the dataset $\{(z_i, \ell_i)\}_{i=1}^{2n} =: \{(x_i, 0)\}_{i=1}^n \cup \{(y_i, 1)\}_{i=1}^n$, then randomly split the equal-sized dataset $\mathcal{D}_{tr}$ and $\mathcal{D}_{te}$, where we train classifier f on $\mathcal{D}_{tr}$ and test on $\mathcal{D}_{te}$. Then the accuracy based on f is $\hat{t}_f = \frac{1}{n}\sum_{(z_i,l_i)\in\mathcal{D}_{te}} \mathbb{I}[f(z_i) = \ell_i]$. Denote the random variable $V_f = \mathbb{I}[f(z) = \ell]$, under null hypothesis $H_0 : \mathbb{P} = \mathbb{Q}$, its expectation is clearly 0.5. Under alternative hypothesis, the expectation is $0.5 + \varepsilon_f$, where ε_f represents the discriminative ability of f. Assume we have m classifiers $f_1, \ldots, f_m$, the test statistics is then defined as $\hat{t} = \frac{1}{m}\sum_j \hat{t}_{f_j}$.

Theorem 1. *Define $e_1 = \frac{1}{m}\sum_j \varepsilon_{f_j}$, $e_2 = \frac{1}{m}\sum_j \varepsilon_{f_j}^2$, $c_0 = \frac{1}{m}\sum_{j\neq k} cov_{H_0}(V_{f_j}, V_{f_k})$, and $c_1 = \frac{1}{m}\sum_{j\neq k} cov_{H_1}(V_{f_j}, V_{f_k})$, where $cov_{H_0}(V_{f_j}, V_{f_k})$ is the covariance between V_{f_j} and V_{f_k} under H_0, and vice versa. Based on the above descriptions, the testing power of the test statistics $\hat{t}$ is asymptotically $\Phi\left(\frac{e_1\sqrt{nm}-\Phi^{-1}(1-\alpha)\sqrt{0.25+c_0}}{\sqrt{0.25-e_2+c_1}}\right)$ with significance level α.*

Proof. Please refer to Appendix A. □

If $m = 1$, the power is reduced to $\Phi\left(\frac{e_1\sqrt{n}-\Phi^{-1}(1-\alpha)/2}{\sqrt{0.25-e_2}}\right)$, If $m > 1$, and every v_{f_j} resulted by the classifier f_j is independent to each other, then $c_0 = c_1 = 0$. The resulted power is $\Phi\left(\frac{e_1\sqrt{mn}-\Phi^{-1}(1-\alpha)/2}{\sqrt{0.25-e_2}}\right)$, which is better than $m = 1$ case. However, it is hard to have independent or uncorrelated f_j in practice. We reduce the covariance by training multiple classifiers with different data split or difference feature partitions [31].

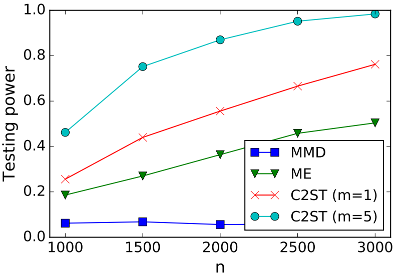


Figure 2: The testing power between different algorithms on Gaussian v.s. Student-*t* distributions.

[1] The probability of rejecting null hypothesis $H_0 : \mathbb{P} = \mathbb{Q}$ when $\mathbb{P} = \mathbb{Q}$.

[2] The probability of accepting null hypothesis $H_0 : \mathbb{P} = \mathbb{Q}$ when $\mathbb{P} \neq \mathbb{Q}$

We demonstrate simple toy data results for empirical justification, where $\mathbb{P}$ is a 50 dimensional standard normal distribution while $\mathbb{Q}$ is the same except for replacing one dimension with a student-t distribution whose degree of freedom is 3. We consider two famous baseline algorithms, maximum mean discrepancy (MMD) [11] and mean embedding (ME) [4, 13]. For C2ST, we train a 2-layer MLP with hidden layer size to be 20. The detailed experiment setting is shown in Appendix B.1. We consider $m = 1$ and $m = 5$ cases. We varies n and report the testing power under $\alpha = 0.05$ significance level. The results are shown in Figure 2. Even though the number of dimensions is only 50, it is already challenging to MMD and ME, which demonstrates that C2ST enjoys the advantage under high dimensional settings. On the other hand, it is clear that increasing number of classifiers significantly boost the testing power even if the classifiers are correlated to each other.

Unmasking as Multiple Two Sample Test. Ionescu et al. [12] employ an unmasking method on the two sets of frames, where they train multiple classifiers by gradually removing the most heavily weighted features from the sample frames, and average the accuracies as an indication of anomaly level. This procedure is essentially a multi-classifier two sample test by training the classifiers with different partition of features, and therefore enjoys the advantages of MC2ST from above. More empirical comparison between strategies of learning multiple classifiers is studied in Section 4.1.

3.3 Increasing Testing Power via Utilizing History

Based on Theorem 1 above, increasing the sample size n improves the testing power. Given a fixed window size, we utilize the history to increase sample size. For each sliding window, we sample a certain number of frames from past time and combine them with the first part of the window as the normal class. Therefore we are comparing the current $t \sim t+n$ frames and b past frames with the second part $t+n+1 \sim t+2n$. To keep the sampling in an online fashion, we maintain a pool of history as the sliding window moves forward, and each time we randomly sample b frames from the pool without replacement. When a frame becomes history, we add it into the history pool with a predefined probability, and replace the previous first frame in the pool.

Screening In practice, the video has considerable variance across the duration of a video. Therefore, when the two consecutive parts are both normal, the history sampling might introduce unnecessary bias to the current windows, which causes the classification accuracy to rise up even though it should be near chance-level. Therefore, we apply a *screening* procedure for the history sampling method: for each window, if the anomaly score is relatively low for the first part then we skip sampling from the history.

3.4 Improving Testing Power via Features

In Theorem 1, $\varepsilon_{f_j} \in [0, 0.5]$ for classifiers can be interpreted as their discriminative ability. Clearly, the larger ε_{f_j} results in stronger testing power. In practice, since we only have limited amount of frames for two windows, we can only adopt simple classifiers (*e.g.* logistic regression) on the extracted features. Therefore, the testing power is implicitly dominated by the quality of the these features.

A commonly used approach for extracting features is to use convolutional neural network pre-trained on large dataset [16, 28, 29] as they require no additional training data. On the other had, many works in video anomaly detection use motion features[7, 12, 21]. Del Giorno et al. [7], Ionescu et al. [12] perform classification on the extracted motion cubes from videos. However, we consider that motion features should embed both the position and amplitude of ongoing motion inside video frames. Also it should allow a reasonable way to perform history sampling introduced in Section 3.3.

We adjust to a more expressive frame level motion descriptor for both better representation and generalization ability that allows us to apply history sampling. Given an input video, we apply the same procedure as Lu et al. [21] to extract 3D gradient features of the consecutive frames. We use the obtained gradient volume of consecutive frames as features, instead of extracting motion cubes. The pixel-level gradient volumes can preserve the spatial information of motion between frames. We also follow Lu et al. [21] to eliminate noise by only preserving the values for regions with motion responses larger than a given threshold. Therefore, we will have a frame-level motion feature with noise filtration.

4 Experiments

We conduct empirical study on four benchmark datasets for detecting abnormal activities, including UCSD pedestrian [23], UMN crowd activity [25], Avenue [21], and the Subway surveillance video [1] datasets. We only use testing videos in each dataset as we do not require normal training data.

We follow Ionescu et al. [12] to use two features to represent the video frames: the appearance feature (pre-trained CNN feature) [28] and motion features [21]. Our improved motion feature is of size 120 × 160 for each frame. We also follow Ionescu et al. [12] to equally partition each frame into 2 × 2 bins to process for better performance. An simple illustration of the idea is shown in Figure 3. Further implementation details can be found is the Appendix B.2.

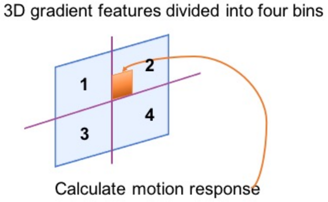

3D gradient features divided into four bins
1
2
3
4
Calculate motion response

Figure 3: An illustration of frame-level motion feature.

Evaluation In practical anomaly detection tasks, we care more about controlling type II errors, i.e. the false negative rate. Therefore, as many works in video anomaly detection[7, 12, 21], we use the area under curve (AUC) to measure performance. Further details are in Appendix C. Note that there are a few videos in datasets, such as Avenue dataset, that only have abnormal frames. This contradicts the assumption of unsupervised learning setting, which assumes the video only contains a few anomalies. So we exclude those special cases for comparison. Since there is no code available of [12], we report our reproduced results of methods in Ionescu et al. [12], which is similar to the reported number in the paper, for fair comparison. For other existing works, we simply use their results as reference.

4.1 Study of Different Ensemble Strategies

In Section 3.2, we analyze the testing power of multiple classifier two-sample test. In this section, we study different strategies for generating multiple classifiers, and show that

AUC(%)	Simple Sampling		Unmasking
/Features	With Replacement	Without Replacement	
Appearance	76.6	76.6	**79.1**
Motion	81.3	78.4	**84.4**

Table 1: Results of different ensemble strategies on Avenue

they can reach comparable results. The detail for training classifiers can be found in Appendix B.2.

The first approach is to equally partition d dimensions to k parts of dimension d/k, and train k classifiers on the sub-features separately. We then average the training accuracy as the anomaly score for each frame. Given that classifiers are trained on different subparts, this procedure can be done in parallel and therefore have the potential to significantly raise the processing speed. An alternative way is to randomly sample a subset of features with replacement for each classifier.

Ionescu et al. [12] instead use *unmasking* to train multiple classifiers given set of frames. Let us suppose the frame features are d dimensional. The unmasking method trains k classifiers and gradually remove m dimensions from the current features at each time based on the classifier weights. The intuition is that if there is a significant difference between the two tested parts, the accuracy will remain high even after removing certain dimensions of features. We performed experiments of these variants on the Avenue dataset and present our results of AUC scores in Table 1. The performance of these different ensemble strategies and the unmasking method shows a trade off between the classifiers' correlation and their discriminative abilities that we discussed in Theorem 1. Compared to the unmasking procedure, the simple ensemble strategies that we observed in this section will reduce the correlation between classifiers to benefit the testing power, since they are trained on vastly different parts of the original features. Also they offer great speed up and computing convenience because of parallel computing. However, the unmasking method has the advantage of high discriminative ability of classifiers despite the overlapping of training data, as they start by training on the whole feature, giving the classifiers more sufficient data. From practical view we can conclude that the unmasking method is a good heuristic, as it has better leveraged result in said trade-off based on the empirical results. Therefore, we stay with this option for our following experiments. The classifier number k and removed feature dimension m is set to 10 and 50.

4.2 Results on Benchmark Datasets

UCSD dataset UCSD dataset contains 2 pedestrian surveillance video sets, Ped1 and Ped2. This dataset is well-labeled with challenging abnormal activities such as riding bikes in crowded roads. UCSD Ped1 has 36 testing videos, which we use 7000 testing frames. UCSD Ped2 has 12 testing videos, which we use 1380 testing frames. Result of our methods are presented in Table 3 and references including supervised methods are in Table 2.

Methods	Ped1(%)	Ped2(%)
Kim and Grauman [14]	59.0	69.3
Mehran et al. [25]	67.5	55.6
Mahadevan et al. [23]	81.8	82.9
Xu et al. [33]	92.1	90.8

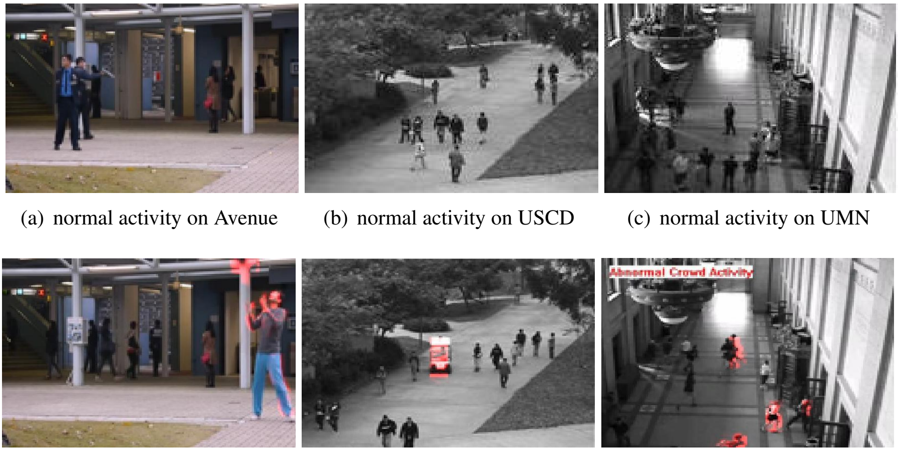
Table 2: Frame AUC of supervised methods on UCSD

We can observe that on appearance feature, the proposed history sampling can indeed improve upon the reproduced unmasking method. Our improved motion feature is also able

AUC(%)	Ped1			Ped2		
/Features	Unmasking	Random Sampling	Screening	Unmasking	Random Sampling	Screening
Appearance	67.9	68.6	**69.0**	80.9	85.0	**87.5**
Improved Motion	71.7	71.2	**71.8**	84.1	**85.8**	85.4

Table 3: Frame AUC on UCSD

to improve performance by giving a better representation. For instance in Ped1 the reproduced result of original motion feature is 67.2%, which is comparable with 67.8% reported in [12], and our improved motion feature is able to bring a 4% improvement. Therefore we stay with this frame-level motion feature that also allows straightforward history sampling. On Ped1 the screened history sampling result on motion feature is at least competitive to the unmasking results if not remarkably better, and on Ped2 they are able to bring larger relative improvements. A demonstration of detection result on several datasets is shown in Figure 4. The localization is visualized via the amplitude of our motion feature.



(a) normal activity on Avenue (b) normal activity on USCD (c) normal activity on UMN

(d) detected anomaly on Avenue (e) detected anomaly on USCD (f) detected anomaly on UMN

Figure 4: Detection results demonstration (best viewed in color)

UMN dataset UMN dataset consists of three main scenes with staged crowd abnormal activities. There are 7739 frames with frame level ground truth. The average result of all three scenes is reported in Table 4. The labeling information can be found in Appendix C.

(a) Existing methods

Methods	AUC(%)
Cong et al. [6]	97.8
Del Giorno et al. [7]	91.0

(b) Our results

AUC(%) / Features	Unmasking	Random Sampling	Screening
Appearance	94.1	94.5	**95.2**
Improved Motion	95.3	94.7	**96.1**

Table 4: Frame AUC on UMN

Although Ionescu et al. [12] already has satisfactory performance on UMN dataset, our methods still demonstrates a stable good performance. A explanation for random sampling not being able to bring further improvement here with the improved motion feature is the possibly biased history. Since there is no constraints to decide whether the sampled history are normal or not, we always have a non-zero chance to sample noisy or anomalous frames. That would introduce undesired bias into our hypothesis test and therefore affect the performance. However, the proposed screening procedure mitigates this problem and bring the performance gain.

Avenue dataset The well-labeled Avenue dataset contains 21 testing videos in which we use 14994 frames, the ground truth is given in pixel level. This dataset is very challenging because it contains various abnormal activities such as throwing objects and running. We report the frame level AUC results of our methods compared them to supervised [21] and unsupervised methods [7, 12] as shown in Table 5.

(a) Existing methods

Methods	AUC(%)
Lu et al. [21]	80.9
Del Giorno et al. [7]	78.3

(b) Our results

AUC(%) / Features	Unmasking	Random Sampling	Screening
Appearance	79.1	79.3	**81.1**
Improved Motion	**84.4**	83.9	83.9

Table 5: Frame AUC on Avenue

The best performance of our methods is brought by improved motion feature. Even considering the two special case videos in Avenue, our improved motion feature alone obtain a score of 82.3%, successfully surpassing the results of other unsupervised and even supervised methods. However, on the motion features, the sampling is not able to further benefit the performance due to the amount of variance in the features. Notice that one of the challenging aspect of Avenue dataset is the camera being closer to the observed crowd, where we will capture more unimportant motions than UCSD and UMN datasets. Also, the activities in the videos have various forms, therefore the motion feature can vary much across time. As we stated before in Section 3.3, the downside of sampling history is the possibility to bring unwanted bias into the two sample test. We calculate the mean variance of motion features on every dimensions in the four bins, and average the results. The avenue dataset has average bin variance of 0.86, while the UCSD Ped2 dataset has average variance of merely 0.19. Under high variance case with small window sizes, limited numbers of history can not benefit the result.

On the other hand, the appearance feature extracted from pretrained CNN is more robust to unimportant motions. Using appearance features, the history sampling with screening has the best result with 2% increment than the original method, showing the effectiveness of adding history frames.

Subway dataset The subway dataset includes two longest surveillance videos of subway exit and entrance gates. The entrance gate video has 144249 frames and the exit gate video has 64901 frames. The labeling information is also in Appendix C. Considering that in this dataset the major abnormal activity with provided ground truth is going in wrong direction and the camera is set facing almost directly to the gate, the appearance feature is not very applicable here, as they are not sensitive to the direction of moving people if they are facing the camera. The results obtained by improved motion feature are presented in Table 6. We report our reproduced result of Ionescu et al. [12] in existing methods considering the possible difference in labeling. We calculate our AUC scores on the whole videos as we do not require any assumption of the occurrence of the abnormal events or training data.

(a) Existing methods

Methods	Entrance gate(%)	Exit gate(%)
Cong et al. [5]	80.0	83.0
Del Giorno et al. [7]	69.1	82.4
Ionescu et al. [12] Motion	69.9	90.0

(b) Our results

AUC(%) / Dataset	Unmasking	Random Sampling	Screening
Entrance gate	**71.7**	71.6	71.6
Exit gate	92.7	92.8	**93.1**

Table 6: Frame AUC on Subway

Our proposed methods of screened history sampling is able to achieve a relative improvement of 3.1% on the exit gate. Similar to the observation from Ionescu et al. [12], our proposed methods has relatively lower scores than supervised method on Entrance gate. This is due to the limitation of general unsupervised methods that cannot rely on labeled training data and will therefore detect some reasonable false positives. Visualized results of such situation can be found in Appendix D. That being said, improvements can still be obtained by our improved motion feature and history sampling.

5 Conclusion

Video anomaly detection has always been a difficult yet interesting task, especially under unsupervised settings. Following Ionescu et al. [12] who use unmasking method to detect sudden changes between consecutive frames inside a sliding window, we offer a new perspective to derive theoretical foundation of this method by connecting it to a statistical analysis method, i.e. the multiple classifier two sample test (MC2ST). Our theoretical analysis is justified by various practical experiments, and gives us directions on how to improve the current method. Based on that, we propose a new frame-level motion feature and the procedure of history sampling, while keeping the online processing style. Our methods have achieved improvement on several surveillance video datasets with theoretical justification.

References

[1] Amit Adam, Ehud Rivlin, Ilan Shimshoni, and Daviv Reinitz. Robust real-time unusual event detection using multiple fixed-location monitors. *TPAMI*, 2008.

[2] Borislav Antić and Björn Ommer. Video parsing for abnormality detection. In *ICCV*, 2011.

[3] Kai-Wen Cheng, Yie-Tarng Chen, and Wen-Hsien Fang. Video anomaly detection and localization using hierarchical feature representation and gaussian process regression. In *CVPR*, 2015.

[4] Kacper P Chwialkowski, Aaditya Ramdas, Dino Sejdinovic, and Arthur Gretton. Fast two-sample testing with analytic representations of probability measures. In *NIPS*, 2015.

[5] Yang Cong, Junsong Yuan, and Ji Liu. Sparse reconstruction cost for abnormal event detection. In *CVPR*, 2011.

[6] Yang Cong, Junsong Yuan, and Ji Liu. Abnormal event detection in crowded scenes using sparse representation. *Pattern Recognition*, 2013.

[7] Allison Del Giorno, J Andrew Bagnell, and Martial Hebert. A discriminative framework for anomaly detection in large videos. In *ECCV*, 2016.

[8] Jayanta Kumar Dutta and Bonny Banerjee. Online detection of abnormal events using incremental coding length. In *AAAI*, 2015.

[9] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. LIBLINEAR: A library for large linear classification. *JMLR*, 2008.

[10] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, 2014.

[11] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *JMLR*, 2012.

[12] Radu Tudor Ionescu, Sorina Smeureanu, Bogdan Alexe, and Marius Popescu. Unmasking the abnormal events in video. In *ICCV*, 2017.

[13] Wittawat Jitkrittum, Zoltán Szabó, Kacper P Chwialkowski, and Arthur Gretton. Interpretable distribution features with maximum testing power. In *NIPS*, 2016.

[14] Jaechul Kim and Kristen Grauman. Observe locally, infer globally: a space-time mrf for detecting abnormal activities with incremental updates. In *CVPR*, 2009.

[15] Andrey Kolmogorov. Sulla determinazione empirica di una lgge di distribuzione. *Inst. Ital. Attuari, Giorn.*, 1933.

[16] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

[17] Sangmin Lee, Hak Gu Kim, and Yong Man Ro. Stan: Spatio-temporal adversarial networks for abnormal event detection. *arXiv preprint arXiv:1804.08381*, 2018.

[18] Chun-Liang Li, Wei-Cheng Chang, Yu Cheng, Yiming Yang, and Barnabás Póczos. Mmd gan: Towards deeper understanding of moment matching network. In *NIPS*, 2017.

[19] Jian Li, Shaogang Gong, and Tao Xiang. Global behaviour inference using probabilistic latent semantic analysis. In *BMVC*, 2008.

[20] David Lopez-Paz and Maxime Oquab. Revisiting classifier two-sample tests. *arXiv preprint arXiv:1610.06545*, 2016.

[21] Cewu Lu, Jianping Shi, and Jiaya Jia. Abnormal event detection at 150 fps in matlab. In *ICCV*, 2013.

[22] Weixin Luo, Wen Liu, and Shenghua Gao. A revisit of sparse coding based anomaly detection in stacked rnn framework. *ICCV, Oct*, 2017.

[23] Vijay Mahadevan, Weixin Li, Viral Bhalodia, and Nuno Vasconcelos. Anomaly detection in crowded scenes. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 1975–1981. IEEE, 2010.

[24] Henry B Mann and Donald R Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, 1947.

[25] Ramin Mehran, Alexis Oyama, and Mubarak Shah. Abnormal crowd behavior detection using social force model. In *CVPR*, 2009.

[26] Huamin Ren, Weifeng Liu, Søren Ingvor Olsen, Sergio Escalera, and Thomas B Moeslund. Unsupervised behavior-specific dictionary learning for abnormal event detection. In *BMVC*, 2015.

[27] Mohammad Sabokrou, Mahmood Fathy, Mojtaba Hoseini, and Reinhard Klette. Real-time anomaly detection and localization in crowded scenes. In *CVPR workshops*, 2015.

[28] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[29] Yan Song, Ian McLoughLin, and Lirong Dai. Deep bottleneck feature for image classification. In *ICMR*, 2015.

[30] Student. The probable error of a mean. *Biometrika*, 1908.

[31] Larry Wasserman. *All of statistics: a concise course in statistical inference*. Springer Science & Business Media, 2013.

[32] Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics bulletin*, 1945.

[33] Dan Xu, Elisa Ricci, Yan Yan, Jingkuan Song, and Nicu Sebe. Learning deep representations of appearance and motion for anomalous event detection. *arXiv preprint arXiv:1510.01553*, 2015.

[34] Wojciech Zaremba, Arthur Gretton, and Matthew Blaschko. B-test: A non-parametric, low variance kernel two-sample test. In *NIPS*, 2013.

[35] Bin Zhao, Li Fei-Fei, and Eric P Xing. Online detection of unusual events in videos via dynamic sparse coding. In *CVPR*, 2011.

A Proof of Theorem 1

Proof. Denote the random variable T for the test statistics $\hat{t}$, then

$$T = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{m}\sum_{j=1}^{m} V_{f_j}^{(i)}\right),$$

where $V_{f_j}^{(i)}$ follows the same distribution but independent to each other for all i. Next, we discuss the property of $V_{f_j}^{(i)}$. To simplify the notation, we use V_{f_j} interchangeably.

Under H_1, V_{f_j} is a Bernoulli random variable with $p = 0.5 + \varepsilon_{f_j}$. Therefore, the mean and variance for $\frac{1}{m}\sum_j V_{f_j}$ are $0.5 + e_1$ and $\frac{0.25-e_2+c_1}{m}$, where $e_1 = \frac{1}{m}\sum_j \varepsilon_{f_j}$, $e_2 = \frac{1}{m}\sum_j \varepsilon_{f_j}^2$, and $c_0 = \frac{1}{m}\sum_{j \neq k} \mathrm{cov}_{H_0}(V_{f_j}, V_{f_k})$. Similarly, under H_0, V_{f_j} is with $p = 0.5$. The mean and variance for $\frac{1}{m}\sum_j V_{f_j}$ Each V_{f_j} is a Bernoulli random variable, with $p = 0.5$ and $p = 0.5 + \varepsilon_{f_j}$ under H_0 and H_1. Therefore, the mean and variance of $\frac{1}{m}\sum_j V_{f_j}$ are 0.5 and $\frac{0.25+c_0}{m}$

By central limit theorem, the test statistics is $T \sim \mathcal{N}(0.5, \frac{0.25+c_0}{mn})$ under H_0 and $T \sim \mathcal{N}(0.5 + e_1, \frac{0.25-e_2+c_1}{mn})$ under H_1. Therefore, given the significance level α, the threshold for accepting or rejecting H_0 is $z_\alpha = 0.5 + \Phi^{-1}(1-\alpha)\sqrt{\frac{0.25+c_0}{mn}}$. We then derive the type-II error as

$$\begin{aligned}
\mathbb{P}_{T \sim \mathcal{N}\left(0.5+e_1, \frac{0.25-e_2+c_1}{mn}\right)}(T < z_\alpha) &= \mathbb{P}_{T \sim \mathcal{N}\left(0, \frac{0.25-e_2+c_1}{mn}\right)}\left(T < \Phi^{-1}(1-\alpha)\sqrt{\frac{0.25+c_0}{mn}} - e_1\right) \\
&= \Phi\left(\sqrt{\frac{mn}{0.25-e_2+c_1}}\left(\Phi^{-1}(1-\alpha)\sqrt{\frac{0.25+c_0}{mn}} - e_1\right)\right) \\
&= \Phi\left(\frac{\Phi^{-1}(1-\alpha)\sqrt{0.25+c_0} - e_1\sqrt{mn}}{\sqrt{0.25-e_2+c_1}}\right)
\end{aligned}$$

Therefor, the power is

$$1 - \Phi\left(\frac{\Phi^{-1}(1-\alpha)\sqrt{0.25+c_0} - e_1\sqrt{mn}}{\sqrt{0.25-e_2+c_1}}\right) = \Phi\left(\frac{e_1\sqrt{mn} - \Phi^{-1}(1-\alpha)\sqrt{0.25+c_0}}{\sqrt{0.25-e_2+c_1}}\right)$$

□

B Implementation details

B.1 Implementation on Toy dataset in Section 3.2

For C2ST in the toy dataset, we train a 2-layer MLP with hidden layer size 20 for 50 epochs. We use RELU as activation function, with the batch size of 16. For optimizer we use the default Adam Optimizer. For $m = 5$ cases, we simply use different random seeds with early stopping for training the classifier.

B.2 Implementation on Benchmark datasets in Section 4

Our frame level motion features have size 120 × 160. We choose different motion thresholds for different datasets to reproduce results in Ionescu et al. [12], and use the same threshold to report results of the proposed algorithm.

For filtering out the noise of the features, we follow Lu et al. [21]. We calculate the motion responses of each region across 5 consecutive frames to obtain a corresponding value that records motion intensity inside each position. Using a threshold, we can find and eliminate features in regions with small motion responses.

They have proved that dividing the frames equally to 2x2 bins and processing them separately can slightly improve the results. Thus we follow some of their settings: the same partitions is done on both our features, for every frame the anomaly score is chosen as the maximum score of the four bins.

The classifiers we use are L2-regularized logistic regression classifier from LIBLINEAR [9]. For history sampling, the size of past frames is set to be 5 or 10, for each individual dataset we use the same history size for all sampling methods. For other hyperparameters, such as stride for the sliding window, we follow Ionescu et al. [12].

C Evaluation

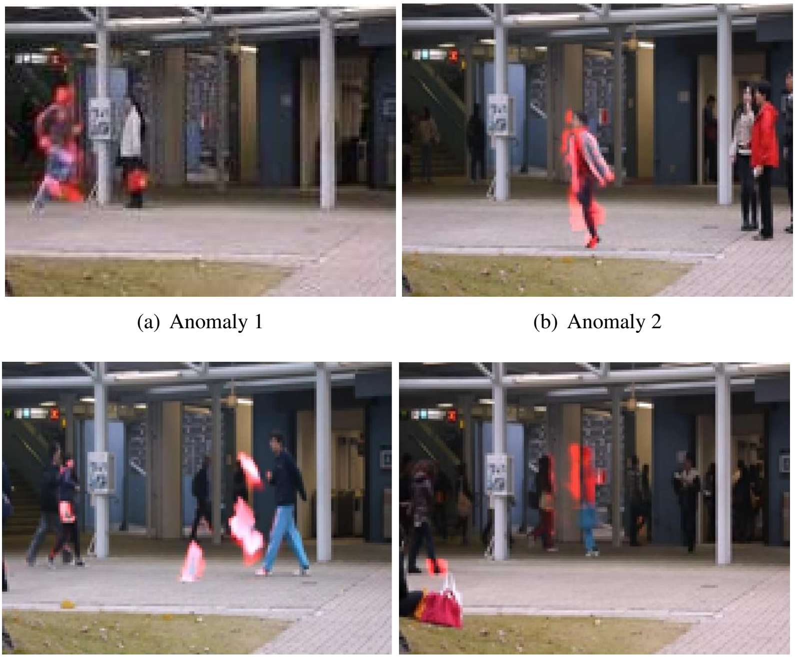
While calculating results on videos, the anomaly scores are smoothed with the same filter as Ionescu et al. [12] before they are used to compute AUC scores with the frame level ground truth. In the setting where pixel level ground truth is provided instead, we consider a frame to be anomalous if it contains abnormal regions.

Also, as the labeling on UMN and Subway datasets are not as precise as Avenue dataset and the UCSD datasets, we manually adjust the labels to be more accurate. For subway dataset, the ground truth is very roughly labeled and only large time intervals are provided for the abnormal activities, so we did the slight adjustment of labeling in reference to the ground truth provided by Adam et al. [1] upon request.

D More Detection Results

The avenue dataset is an interesting case as there are various forms of abnormal activities presented. Our improved motion feature demonstrate a great fit for detecting the anomalies. The visualized results are shown in Figure 5.

The subway dataset has a more complex environment, so as mentioned previously in the experiments, there are some reasonable false positive detections due to the limitation of unsupervised methods. Here we observe a simple case of true positive(wrong direction) and false positive(people jogging to the subway) detections by improved motion feature in Figure 6.



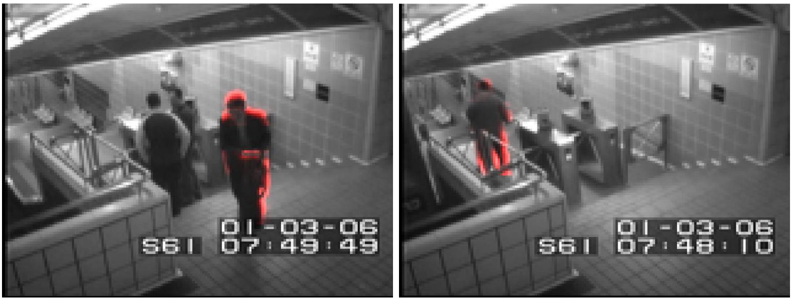
(a) Anomaly 1

(b) Anomaly 2

(c) Anomaly 3

(d) Anomaly 4

Figure 5: More detection results on Avenue dataset



(a) True Positive

(b) False Positive

Figure 6: True positive and false positive detections on Subway dataset