Temporal Cues from Socially Unacceptable Trajectories for Anomaly Detection

Madan, Neelu; Farkhondeh, Arya; Nasrollahi, Kamal; Escalera, Sergio; Moeslund, Thomas B.

Published in:
2021 IEEE/CVF International Conference on Computer Vision Workshops, ICCVW 2021

DOI (link to publication from Publisher):
10.1109/ICCVW54120.2021.00244

Publication date:
2021

Document Version
Accepted author manuscript, peer reviewed version

Temporal Cues from Socially Unacceptable Trajectories for Anomaly Detection

Neelu Madan
Aalborg University
nema@create.aau.dk

Arya Farkhondeh
Sapienza University of Rome
arya.farkhondeh@gmail.com

Kamal Nasrollahi
Aalborg University and Milestone Systems A/S
kn@create.aau.dk

Sergio Escalera
Universitat de Barcelona, Computer Vision Center, and Aalborg University
sergio@maia.ub.es

Thomas B. Moeslund
Aalborg University
kn@create.aau.dk

Abstract

State-of-the-Art (SoTA) deep learning-based approaches to detect anomalies in surveillance videos utilize limited temporal information, including basic information from motion, e.g., optical flow computed between consecutive frames. In this paper, we compliment the SoTA methods by including long-range dependencies from trajectories for anomaly detection. To achieve that, we first created trajectories by running a tracker on two SoTA datasets, namely Avenue and Shanghai-Tech. We propose a prediction-based anomaly detection method using trajectories based on Social GANs, also called in this paper as temporal-based anomaly detection. Then, we hypothesize that late fusion of the result of this temporal-based anomaly detection system with spatial-based anomaly detection systems produces SoTA results. We verify this hypothesis on two spatial-based anomaly detection systems. We show that both cases produce results better than baseline spatial-based systems, indicating the usefulness of the temporal information coming from the trajectories for anomaly detection. We observe that the proposed approach depicts the maximum improvement in micro-level Area-Under-the-Curve (AUC) by 4.1% on CUHK Avenue and 3.4% on Shanghai-Tech over one of the baseline method. We also show a high performance on cross-data evaluation, where we learn the weights to combine spatial and temporal information on Shanghai-Tech and perform evaluation on CUHK Avenue and vice-versa.

1. Introduction

Video anomaly detection is a sub-domain of behavior understanding, where anomalies for applications such as theft detection, traffic light jumping, and fighting, etc. are getting increasingly relevant with the accessibility and proliferation of video surveillance. There are multiple challenges associated with anomaly detection including the vague definition of anomalous behavior, i.e., anomaly changes with the context. An example to illustrate the context can be that driving a vehicle on a pedestrian street is considered anomalous while it is normal in the context of a road. Additionally, by definition anomalies are rare to anticipate, which consequently leads to the failure of supervised learning methods due to imbalanced datasets.

Therefore, unsupervised and weakly supervised anomaly detection approaches have recently gained interest. Common examples are reconstruction [13] and prediction [19] based anomaly detection. Reconstruction-based anomaly detection systems reconstruct the current frame and prediction-based ones predict the future frame. If the reconstruction/prediction error is low, the current/future frame is normal, otherwise abnormal. State-of-the-Art deep learning approaches for anomaly detection are only trained for normal events, with the hypothesis that the reconstruction/prediction error for anomalous frames is high. However, neural networks sometimes learn to reconstruct/predict even anomalous frames with low errors. This reduces the discriminative power of the neural network to classify a frame as abnormal or normal. To over-

come this drawback, memory-augmented auto-encoders [30, 10] are proposed. The memory-augmented auto-encoders [30, 10] contain an extra memory module along with a prediction/reconstruction-based network. The memory module learns to cluster the normal events in the training data and finally uses a one-class classification approach to identify the anomalies. It basically creates a prototype for each normal event in the training data and prevents the network from generalizing for abnormal events. Despite the great achievements of SoTA methods in anomaly detection, still, there is room for improvements. SoTA approaches are mostly using spatial information for anomaly detection and utilizing temporal information has been limited to gradient or optical flow computed between consecutive frames. Obtaining the optical flow for large datasets is a time-consuming and computationally expensive process. This is the reason that most anomaly detection systems utilizing optical flow extract this from only two frames [9]. The object's trajectories, which implicitly include the history of motion [31] are better choices and are also computationally efficient. However, contextual anomalies such as walking in restricted zones and behavioral anomalies such as dancing or jumping are not captured by using only trajectories. Therefore, we need an appropriate balance of spatial and temporal information for robust anomaly detection.

In this paper, we hypothesize that fusing temporal anomaly detection scores (based on trajectories) with spatial anomaly detection scores (based on SoTA methods) increases the accuracy of these systems, regardless of the network architecture used for spatial anomaly detection. To encode the long-range dependencies for the video anomaly detection, we use these trajectories to detect the anomalies using our proposed temporal network based on Social Generative Adversarial Networks (Social GANs) [12]. We implicitly consider social interaction among different objects in the scene during anomaly detection using trajectories because of the presence of social pooling layer in Social GANs [12]. We verify our hypothesis by using different baselines, i.e., prediction-based system of Liu *et al*. [19] and memory-based system of Park *et al*. [30] for our spatial network. The prediction-based system of Liu *et al*. [19] predicts a future frame from the past four frames by minimizing intensity, gradient, and flow loss. However, the memory-based system of Park *et al*. [30] incorporates additional memory modules for both prediction-based and reconstruction-based anomaly detection. For the inclusion of temporal information from trajectories, we learn a score level fusion of anomaly detection scores obtained from the temporal and spatial networks.

We verify that there is improvement in frame-level AUC (a commonly used metric for video anomaly detection) for each baseline by using the complementary information from trajectories. There is an improvement of 1.7% on CUHK Avenue [21] and 1.8% on Shanghai-Tech [19] for Liu *et al*. [19]. The inclusion of trajectories in Park *et al*. [30] shows an improvement of 4.1% on CUHK Avenue [21] and 3.3% on Shanghai-Tech [19] for reconstruction-based and an improvement of 0.1% on CUHK Avenue [21] and 3.3% on Shanghai-Tech [19] for prediction-based approaches. We also perform some additional experiments on cross-database generalization, where we learn parameters on Shanghai-Tech [19] and use them to evaluate CUHK Avenue [21] or vice-versa. We observe an overall increase in performance even in-case of cross-databases experiments, i.e., from CUHK Avenue [21] to Shanghai-Tech [19] have an improvement of 0.7% and from Shanghai-Tech [19] to CUHK Avenue [21] have an improvement of 1.8% in the AUC over the baseline by Li *et al*. [19]. The late fusion of spatial and temporal information makes our approach applicable to any SoTA anomaly detection method.

2. Related Work

Systems to deal with the task of video anomaly detection are getting complex with the evolution of complex anomalies and new datasets. The methods use for video anomaly detections are broadly classified into two categories namely spatial-based and temporal-based anomaly detection.

2.1. Anomaly Detection Using Spatial Cues

Anomaly detection systems utilizing spatial information can be further classified into four sub-categories: Reconstruction, Prediction, Hybrid and Object-centric approaches. Reconstruction-based approaches seek to learn normalcy, where the expectation is that anomalous activity will have a large reconstruction error, comparing the input with its reconstruction. This approach has shown promise due to the era of deep learning and specifically the convolutional autoencoder (CAE) and the generative adversarial network (GAN) [11]. The work of Hasan *et al*. [13] is the first example of applying CAE and comparing it to hand-crafted features like Histograms of Oriented Gradients (HOG) and Histograms of Optical Flows (HOF), showing the potential of learned representations. Similar approach is seen using GANs [32, 29]. Prediction-based approaches argue that anomalous actions are naturally harder to predict. This approach is pioneered by Liu *et al*. [19], using a sliding time window to predict the future frame. The future prediction is then compared to the actual input. This is further expanded by Rodrigues *et al*. [33] using multiple timescales. Hybrid approaches [37] [27] [34] are combining both the reconstruction and prediction aspects. To avail the success of deep learning-based object-detection, few anomaly detection approaches such as [15, 9, 8] incorporates anomaly score based on object detection rather than on frame-level.

Training unsupervised methods for a complex task such as anomaly detection is challenging due to limited guidance

during learning, compared to supervised learning. There are some methods that are adding some prior information to the above approaches for improving accuracy. A common approach to aid in the learning is to use pre-trained systems to impose what is already known and learned, either in the form of optical flow [9], object detectors [32, 15, 36], skeletons [27], or memory augmentation [10, 30]. The downside of many of these methods is the limited use of contextual information. In recent years, memory-augmentation networks that are using external memory to extend the capabilities of the neural network are used, e.g., Gong. *et al*. [10] proposed a memory-augmented deep autoencoder, where rather than reconstructing the frame directly, the representation obtained from the encoder part is used for querying the most relevant information out of the memory for reconstruction. These types of networks mitigated the issue that abnormal frames can also be reconstructed with a small error.

2.2. Anomaly Detection Using Temporal Cues

SoTA anomaly detection approaches are mostly using spatial cues, while taking only limited temporal information into consideration. For example, Liu *et al*. [19] use optical flow between consecutive frames, Ionescu *et al*. [15] use backward gradient between the previous and current frame and forward gradient between current and next frame. Later, Georgescu *et al*. [9] verified that optical flow is better to capture motion in the context of anomaly detection, so they replaced forward and backward gradient in Ionescu *et al*. [15] by forward and backward optical flow.

There are limited approaches such as Morias *et al*. [27] and Rodrigue *et al*. [33] including trajectory for anomaly detection. Morias *et al*. [27] uses a skeleton-based representation of trajectories, which needs additional annotations for gaits in human body. To further expand this work, Rodrigue *et al*. [33] also uses pose-based trajectories but extracted features at multiple scales. The limitation of posed-based trajectories is that they are only applicable for human anomalies, and non-human anomalies such as vehicles on the pedestrian street or unattended luggage cannot be detected.

Some examples of anomaly detection using trajectories on traffic and old datasets include [3, 4], which are based on the clustering of trajectories using hand-crafted features and distance measures between the trajectories. In this case, the clusters with small support are anomalous. Some other statistical approaches use for anomaly detection include probabilistic modeling and learning of normal trajectories, e.g., [28] applied Hidden Markov Model followed by K-Mean clustering. A rule-based classifier implemented by [18] applies different rules at multiple granularities to classify each data-point as normal or abnormal. In [17], a Bayesian network is used to model the underlying distribution. Some initial deep learning-based approaches such as [24], and [35] still rely on designing the input features in the training set. Some years later, more sophisticated approaches such as using a fully automated LSTM auto-encoder are proposed [2, 16]. Approach by Bouritsas *et al*. [2] and Ji *et al*. are applicable even for non-human anomalies, but they are not performing well on large scale anomaly datasets such as Shanghai-Tech [19]. These methods do not include any social interaction for anomaly detection using trajectories.

There exist some research using social interaction for trajectory prediction. Some examples are Gupta *et al*. [12] and Alahi *et al*. [1]. The basic architecture of both approaches includes a single LSTM for each trajectory followed by a social pooling layer to model the interaction. Social GAN [12] however encouraged diverse prediction by including variety loss, which leads to the prediction of near to real trajectories. In this paper, we propose a novel method for prediction-based anomaly detection using trajectories. Our architecture is mainly motivated by Social GANs [12], where we classify the socially possible trajectories to normal or abnormal based on their prediction error. We then show that this prediction-based anomaly detection system utilizing temporal information in form of trajectories can complement spatial-based anomaly detection sytems, resulting in SoTA performance on two benchmark datasets. To the best of our knowledge, none of the previous works for anomaly detection on surveillance datasets explored the inclusion of socially acceptable trajectories generated via tracker as an additional cue.

3. Proposed System

The block diagram of the proposed system is shown in Figure 1. The main idea of our proposed approach is to utilize social interaction embedded in trajectories to develop a temporal-based anomaly detection system and then use that to complement SoTA spatial-based anomaly detection systems. To achieve this, our proposed system contains two branches, i.e., the spatial branch, which detects the anomalies by mostly using image features, and the temporal branch, which detects the anomalies using trajectories.
In this paper, we use two different SoTA methods for our spatial branch, i.e., Liu *et al*. [19] and Park *et al*. [30], which produce spatial anomaly detection scores. Section 3.2 contains a detailed description of the spatial baseline methods used in our approach. The input to the temporal branch are trajectories, obtained by running tracker [25] on CUHK Avenue [21] and Shanghai-Tech [19]. The generated trajectories are provided as input to the prediction-based anomaly detection network, which also incorporates the features involved with social interaction among the different trajectories. The proposed prediction-based anomaly detection is based on Social GANs [12] and is described in section 3.1. Once we have anomaly score estimated from both spa-

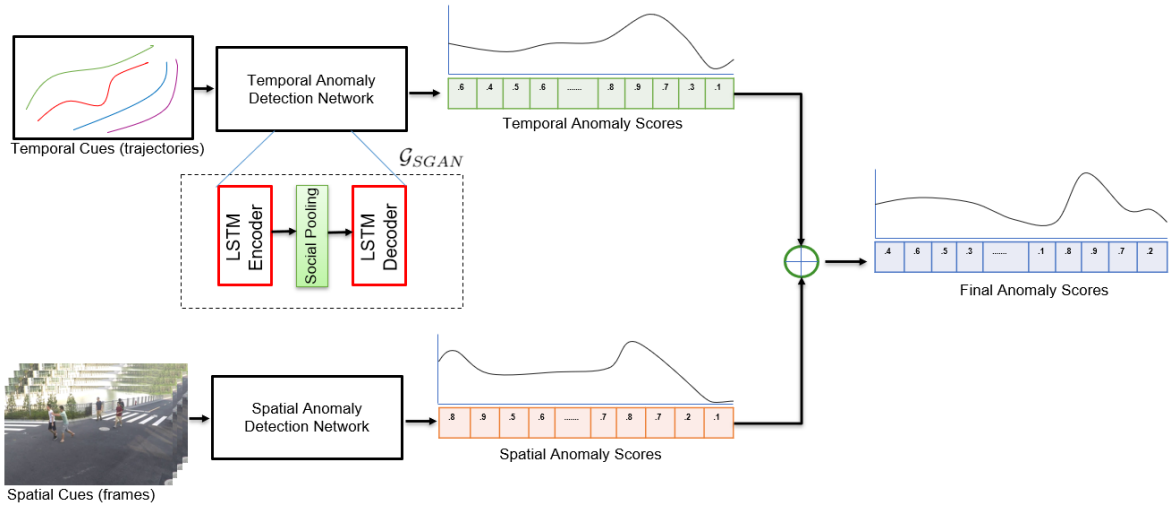

Figure 1. Proposed system for anomaly detection. It contains spatial and temporal branch, respectively. Weighted combination of spatial and temporal anomaly detection is used to generate the final anomaly score.

tial and temporal branches, a weighted score-level fusion is performed to generate the final scores.

3.1. Temporal Branch

We propose a method based on Social GAN [12] to detect anomalous trajectories. The generator ($\mathcal{G}_{SGAN}$) network is an LSTM based encoder-decoder, where one LSTM is used for predicting a single trajectory. The prediction of human trajectories in a crowded scene also depends on social interaction among different human beings. Therefore, $\mathcal{G}_{SGAN}$ contains a social pooling module to encode this interaction. The discriminator ($\mathcal{D}_{SGAN}$) is a LSTM encoder network that classifies the output trajectories as real or fake and encourages the generator to predict socially possible trajectories.

The input to the generator ($\mathcal{G}_{SGAN}$) network is a fixed number of past tracklets from the generated trajectories, which in turn further generates a fixed number of future tracklets. Attention-gated tracker [25] is used to generate the trajectories on CUHK Avenue [21] and Shanghai-Tech [19] datasets. The objective function used for predicting future trajectory is the combination of average displacement error (ADE), final displacement error (FDE), and variety loss. ADE is computed as l_2 distance between the predicted and actual points in the future trajectory, FDE is the deviation in the final position with respect to ground-truth (GT), and variety loss is added to mitigate the redundancy in the predicted trajectories. To transform the trajectory prediction network for anomaly detection, i.e., detecting socially unacceptable trajectories, we compute the total error (TE) by combining ADE and FDE for each tracklet. The tracklet is finally classified as normal or anomalous based on the Total Error (TE), which is also called here as temporal anomaly detection score:

$$TE(t) = ADE(t) + FDE(t), \tag{1}$$

$$S_{sgan}(t) = \frac{TE(T_t, \hat{T}_t) - \min_t TE(T_t, \hat{T}_t)}{\max_t TE(T_t, \hat{T}_t) - \min_t TE(T_t, \hat{T}_t)}, \tag{2}$$

where, T and $\hat{T}$ are actual and predicted trajectory, respectively, and $S_{sgan}(t)$ is the normalized score obtained from social GANs for each tracklet t. We later combine the normalcy scores from temporal and spatial branches. Therefore, we update the total error ($S_{sgan}(t)$) obtained from social GAN (Equation 2) to obtain the normalcy score ($S_{temporal}$), which is also called the temporal network output in this work:

$$S_{temporal}(t) = (1 - S_{sgan}(t)), \tag{3}$$

3.2. Spatial Branch

To show that the proposed temporal-based anomaly detection system using trajectories can improve the performance of different spatial-based anomaly detection systems, we use two different networks in different experiments in the spatial branch of our proposed system. These are future frame prediction-based by Liu *et al*. [19] and memory-based reconstruction/prediction by Park *et al*. [30]. The prediction-based by Liu *et al*. [19] proposed a GAN based method, where generator network aims to generate realistic future frames and discriminator module aims to discriminate between real and generated future frames. Finally, the generated future frame is classified as abnormal or

normal based on its quality. The generated normal frames have better quality in comparison to the abnormal frames. This network uses minimal temporal information in the form of optical flow between consecutive frames and optimizes for intensity, gradient, and flow loss. The memory-augmented anomaly detection by Park *et al*. [30] contains an additional memory module which records prototypical pattern of normal data. The memory module is included with both prediction and reconstruction based anomaly detection networks. Park et al. [30] uses convolutional auto-encoders for both reconstruction and prediction networks. It optimizes both prediction/reconstruction auto-encoders by minimizing prediction/reconstruction, compactness, and separateness loss. The compactness loss encourages the query to the nearest item in the memory and the separateness loss encourages the discriminative power of the memory items. Peak Signal to Noise Ratio (PSNR) by Mathieu *et al*. [26], a commonly used method for image quality assessment, is used for evaluating the predicted/reconstructed frames in both cases:

$$PSNR(I, \hat{I}) = 10 \log_{10} \frac{[\max_{\hat{I}}]^2}{\frac{1}{N} \sum_{i=0}^{N} (I_i - \hat{I}_i)^2}, \quad (4)$$

where, I is actual and $\hat{I}$ is predicted/reconstructed frame. Higher PSNR of the predicted frame increases the probability of it being normal. Then the PSNR score calculated for each frame in a video to generate the spatial anomaly detection score (5) [19]:

$$S_{spatial}(t) = \frac{PSNR(I_t, \hat{I}_t) - \min_t PSNR(I_t, \hat{I}_t)}{\max_t PSNR(I_t, \hat{I}_t) - \min_t PSNR(I_t, \hat{I}_t)}, \quad (5)$$

where, $S_{spatial}(t)$ is the normalized score for t_{th} frame, I_t and $\hat{I}_t$ are actual and predicted/reconstructed frame, respectively, for tracklet t. .

3.3. Parameter Learning

We propose a parameter learning approach to fuse the information from spatial branch and temporal branch at the score level. Thus, we learn two parameters, one for each score vector. The fusion is defined as follows:

$$S_{Total}(t) = \mathcal{F}(\alpha S_{spatial}(t) + \beta S_{temporal}(t)), \quad (6)$$

where α and β are the parameters that we learn to weigh the contribution of spatial network output ($S_{spatial}$) and temporal network output ($S_{temporal}$), respectively. $\mathcal{F}$ is the activation function which is Sigmoid in our case. To form the learning problem, we minimize the binary cross-entropy loss function.

4. Experiments and Results

This section contains details of the evaluation metric, datasets and implementation used in our experiments. The later part of this section also contains quantitative and qualitative results documenting the performance of the introduced temporal-based anomaly detection system using the socially unacceptable trajectories, and its contribution to the proposed system when used with spatial-based anomaly detection systems.

4.1. Evaluation Metrics

The proposed system is evaluated using Receiver Operation Characteristic (ROC) [6] obtained by changing the normality threshold, i.e., fused scores obtained from spatial and temporal network in our case. Area Under the Curve (AUC) is a cumulative measure of accuracy for all possible normality thresholds and used for the accuracy evaluation. A higher value of AUC indicates a better system.

4.2. Datasets

We used two publicly available datasets namely CUHK Avenue [21] and Shanghai-Tech [19] for the training of the baseline models. CUHK Avenue [21] contains 16 training and 21 testing videos with a total of 47 anomalous events. The anomalous events in this dataset are loitering, running, and throwing objects. Shanghai-Tech [19] contains 330 training and 107 test videos with 130 abnormal events. The anomalous events are snatching, chasing, running, fighting, cyclist and vehicles on pedestrian street.

To train the temporal anomaly detection network, trajectory datasets are generated by providing training and testing images from CUHK Avenue [21] and Shanghai-Tech [19] to the attention-gated tracker of Madan *et al*. [25]. The tracking results contain the coordinates of the bounding box along with the object (Identification) ID. The obtained results are converted to a trajectory dataset by converting bounding box coordinates to the center location. Each center position along with the associated ID represents a single tracklet. Object positions associated with the same ID are joined together to form a single trajectory.

4.3. Training and Testing the Proposed System

The baseline architectures of [30, 19] are trained for 15 epochs each on Nvidia RTX 2080 Ti GPU on Shanghai-Tech [19] dataset, which took $\sim$12hrs to complete. We use pre-trained models for CUHK Avenue [21] dataset. Temporal network is trained for 200 epochs individually for each dataset with a batch size of 64 on Nvidia RTX 2080 Ti GPU, which took $\sim$2hrs to complete the training.

At the testing time, we obtain the score vectors from each spatial and temporal branch of our proposed system, which are provided as input to our paramater learning scheme. The

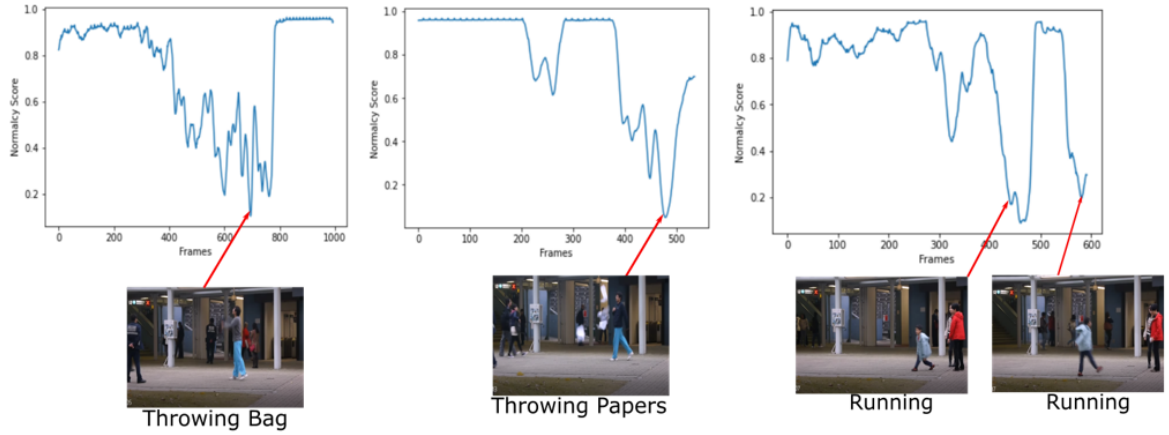

Figure 2. Illustrating the anomalies detected by our strategy on avenue dataset. This includes mostly individual anomalies such as throwing bag (left), throwing paper (middle) and running (right).

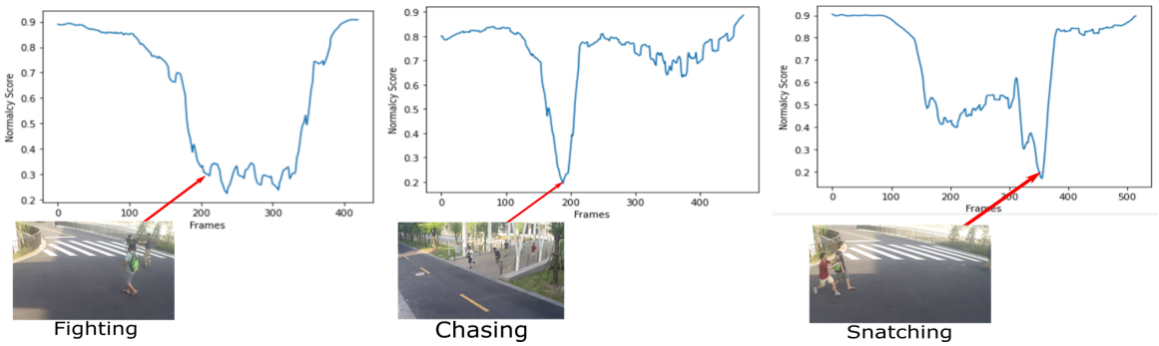

Figure 3. Illustrating the anomalies detected by our strategy on Shanghai-Tech [19] dataset, involving social interaction such as fighting (left), chasing (middle), and snatching (right).

Method	CUHK Avenue(%)	Shanghai-Tech(%)
Hasan *et al.* [14]	80.0	60.9
Del *et al.* [5]	78.3	-
Luo *et al.* [22]	77.0	-
Hinami *et al.* [22]	80.9	-
Lu *et al.* [22]	80.9	-
Ionescu *et al.* [23]	80.6	-
Luo *et al.* [23]	81.7	68.0
Liu *et al.* [20]	84.4	-
Ours (Temporal Only: SGAN)	65.0	69.7
Spatial Only: Liu et. al. [19]	85.1	72.8
Ours (Spatial: Liu et. al., Temporal: SGAN)	**86.8**	**74.6**
Spatial Only: Park et. al. - Pred [30]	88.5	70.5
Ours (Spatial: Park et. al. - Pred., Temporal: SGAN)	**88.6**	**73.8**
Spatial Only: Park et. al. - Reconst [30]	82.8	69.8
Ours (Spatial: Park et. al. - Reconst., Temporal: SGAN)	**86.9**	**73.2**

Table 1. Comparing the frame-level AUC score (in %) of the proposed system with the SoTA approaches and their corresponding spatial anomaly detection branch. Higher frame-level AUC indicate the better performance.

learned parameters are used to weigh the spatial and temporal anomaly scores to generate the final scores. We performed micro-level evaluation, as done in [9, 30], where we concatenate all the sequence and learned the parameters for

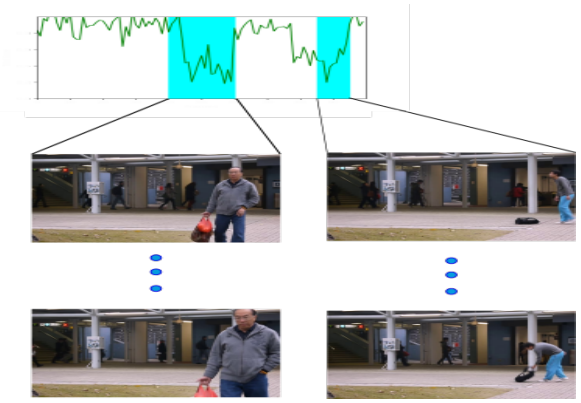

Figure 4. An example of an anomaly sequence "walking in wrong direction and throwing bag", from sequence 6 in CUHK Avenue [19], is not detected by the baseline method but it is detected when complemented with trajectory information using the proposed system.

the entire dataset.

4.4. Qualitative results

Figure 2 and 3 illustrate visual results on CUHK Avenue [21] and Shanghai-Tech [19] datasets. CUHK Avenue [21] mostly contains individual anomalies, which includes limited social interaction, but our proposed combination still improved the anomaly detection by considering individual trajectories. On the other hand, anomalies in Shanghai-Tech [19] involve small groups interaction such as snatching, fighting. Figure 3 depicts that the proposed combination detected anomalies like fighting, chasing, and snatching, all of which involve interaction between two people. Thus, our method improves anomaly detection not only in the case of social interaction, but also involving individual trajectories.

As an illustration of a corrected case, Figure 4 shows an anomaly corresponding to a person moving back and forth to pick-up the bag. This anomaly remains undetected by the baseline method, i.e., Liu *et al*. [19]. However, it is detected by the proposed system. The reason is that continuous back and forth motion is considered as an unacceptable social trajectory.

4.5. Quantitative Results

As depicted in Table 1, the AUC score on CUHK Avenue [21] and Shanghai-Tech [19] using only temporal branch are 65.0% and 69.7%, respectively. It can be observed from these results that trajectories alone are unable to generate competitive results against SoTA methods. The trajectories used in our experiments are constructed using center point, which do not contain much information about the spatial and appearance features of the different objects. Therefore, anomaly detection by simply using these trajectories generate lower AUC scores compared to SoTA. However, when fused with spatial information, as illustrated in Figure 1, temporal information generated by socially acceptable trajectories contributes in increasing the performance of SoTA spatial-based anomaly detection systems by a large margin, as shown in Table 1.

It can be observed from the results shown in Table 1 that the proposed system outperforms listed SoTA approaches including our baseline architecture by Liu *et al*. [19] on CUHK Avenue [21] by 3.4% and on Shanghai-Tech [19] by 1.8%. It outperforms the baseline architecture by Park *et al*. [30] in both forms of 1) reconstruction-based: CUHK Avenue [21] by 4.1% and Shanghai-Tech [19] by 3.3% and 2) prediction-based: CUHK Avenue [21] by 0.1% and Shanghai-Tech [19] by 3.4%. It can be observed from Table 1 that the information from trajectories is complimenting the baseline architectures irrespective of the underlying network architecture in the spatial branch of our proposed system. We didn't compare our results against other SoTA approaches, like [32, 15, 36] in this table as they use additional prior knowledge in form of object-detection, which could be included in our system as future work.

Furthermore, the proposed approach does not optimize the feature space with any additional supervision. Some approaches such as Geogescu *et al*. [9] and Feng *et al*. [7] use additional supervision with pseudo labels to improve the latent features, enhancing accuracy of anomaly detection. On the other hand, our approach learns an accurate fusion of temporal and spatial scores without modifying the underlying feature space through additional supervision. Weakly supervised approach by Geogescu *et al*. [9] has an AUC of 92.3% on CUHK Avenue [21] and 82.7% on Shanghai-Tech [19]. Weakly supervised approach of Feng *et al*. [7] has an AUC of 94.3% on Shanghai-Tech [19]. Comparing with weakly supervised approaches, we observed that our approach has competitive results while having less supervision.

4.6. Cross-data Evaluation Results

We also verified that learning parameters on a source dataset and testing them on a target dataset with similar anomalies also improves the overall score. We used prediction-based anomaly detection by Liu *et al*. as the baseline for this experiment. It can be observed from Table 2 that the AUC on Shanghai-Tech [19], i.e., 73.1% is better than the baseline, i.e., 72.4% by 0.7% and CUHK Avenue [21], i.e., 86.9% is better than baseline, i.e., 85.1% by 1.8%.

Baseline		Proposed	
Dataset	AUC(%)	Train → Test	AUC(%)
CUHK Avenue	85.1	Shanghai-Tech → CUHK Avenue	86.9
Shanghai-Tech	72.4	CUHK Avenue → ShanghaiTech	73.1

Table 2. Cross-data experiments depicting that the learned parameters on one dataset improves the scores on another.

5. Conclusion

In this paper we hypothesize that temporal information obtained from socially unacceptable trajectories can be used for developing a temporal-based anomaly detection system. Then, we further hypothesize that such a temporal-based anomaly detection system can contribute to improving the performance of SoTA spatial-based anomaly detection systems. To verify these, we propose a system with two branches (one for the spatial and one for the temporal domain) that fuses the results of the two domains at score level. We verify that socially unacceptable trajectories provide discriminative information to identify anomalies in real world surveillance datasets, for two different spatial-based systems employed in the spatial branch of our system. We plan as future work to evaluate different temporal and spatial anomaly detection models in both branches of the proposed scheme and analyze for their complementarity. We also plan to incorporate the prior knowledge from object detection or skeleton for anomaly detection.

6. Acknowledgements

This work was supported by the Milestone Research Programme at Aalborg University (MRPA), by the Spanish project PID2019-105093GB-I00 (MINECO/FEDER, UE), and by ICREA under the ICREA Academia programme.

References

[1] Alexandre Alahi, Kratarth Goel, Vignesh Ramanathan, Alexandre Robicquet, Li Fei-Fei, and Silvio Savarese. Social lstm: Human trajectory prediction in crowded spaces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[2] G. Bouritsas, S. Daveas, A. Danelakis, and S. C. A. Thomopoulos. Automated real-time anomaly detection in human trajectories using sequence to sequence networks. In *2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–8, 2019.

[3] Yingyi Bu, Lei Chen, Ada Wai-Chee Fu, and Dawei Liu. Efficient anomaly monitoring over moving object trajectory streams. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09, page 159–168, New York, NY, USA, 2009. Association for Computing Machinery.

[4] Chao CHEN, Daqing Zhang, Pablo Samuel CASTRO, Nan Li, Lin Sun, Shijian LI, and Zonghui WANG. iBOAT : isolation-based online anomalous trajectory detection. *IEEE Transactions on Intelligent Transportation Systems*, 14(2):806–818, June 2013.

[5] Allison Del Giorno, J. Andrew Bagnell, and Martial Hebert. A discriminative framework for anomaly detection in large videos. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 334–349, Cham, 2016. Springer International Publishing.

[6] Tom Fawcett. An introduction to roc analysis. *Pattern Recognit. Lett.*, 27(8):861–874, 2006.

[7] Jia-Chang Feng, Fa-Ting Hong, and Wei-Shi Zheng. Mist: Multiple instance self-training framework for video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14009–14018, June 2021.

[8] Mariana-Iuliana Georgescu, Antonio Barbalau, Radu Tudor Ionescu, Fahad Shahbaz Khan, Marius Popescu, and Mubarak Shah. Anomaly detection in video via self-supervised and multi-task learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12742–12752, June 2021.

[9] Mariana Iuliana Georgescu, Radu Ionescu, Fahad Shahbaz Khan, Marius Popescu, and Mubarak Shah. A background-agnostic framework with adversarial training for abnormal event detection in video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2021.

[10] Dong Gong, Lingqiao Liu, Vuong Le, Budhaditya Saha, Moussa Reda Mansour, Svetha Venkatesh, and Anton van den Hengel. Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In *IEEE International Conference on Computer Vision (ICCV)*, 2019.

[11] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. Book in preparation for MIT Press.

[12] Agrim Gupta, Justin Johnson, Li Fei-Fei, Silvio Savarese, and Alexandre Alahi. Social gan: Socially acceptable trajectories with generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[13] M. Hasan, J. Choi, J. Neumann, A. K. Roy-Chowdhury, and L. S. Davis. Learning temporal regularity in video sequences. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 733–742, 2016.

[14] Mahmudul Hasan, Jonghyun Choi, Jan Neumann, Amit K. Roy-Chowdhury, and Larry S. Davis. Learning temporal regularity in video sequences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[15] Radu Tudor Ionescu, Fahad Shahbaz Khan, Mariana-Iuliana Georgescu, and Ling Shao. Object-centric auto-encoders and dummy anomalies for abnormal event detection in video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[16] Yufan Ji, Lunwen Wang, Weilu Wu, Hao Shao, and Yanqing Feng. A method for lstm-based trajectory modeling and abnormal trajectory detection. *IEEE Access*, 8:104063–104073, 2020.

[17] F. Johansson and G. Falkman. Detection of vessel anomalies - a bayesian network approach. In *2007 3rd International Conference on Intelligent Sensors, Sensor Networks and Information*, pages 395–400, 2007.

[18] Xiaolei Li, Jiawei Han, Sangkyum Kim, and Hector Gonzalez. ROAM: rule- and motif-based anomaly detection in massive moving object data sets. In *Proceedings of the Seventh SIAM International Conference on Data Mining, April*

26-28, 2007, Minneapolis, Minnesota, USA, pages 273–284. SIAM, 2007.

[19] W. Liu, W. Luo, D. Lian, and S. Gao. Future frame prediction for anomaly detection - a new baseline. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6536–6545, 2018.

[20] Yusha Liu, Chun-Liang Li, and Barnabás Póczos. Classifier two sample test for video anomaly detections. In *British Machine Vision Conference 2018, BMVC 2018, Newcastle, UK, September 3-6, 2018*, page 71. BMVA Press, 2018.

[21] Cewu Lu, Jianping Shi, and Jiaya Jia. Abnormal event detection at 150 fps in matlab. In *IEEE International Conference on Computer Vision (ICCV)*, 2013.

[22] C. Lu, J. Shi, and J. Jia. Abnormal event detection at 150 fps in matlab. In *2013 IEEE International Conference on Computer Vision*, pages 2720–2727, 2013.

[23] Weixin Luo, Wen Liu, and Shenghua Gao. A revisit of sparse coding based anomaly detection in stacked rnn framework. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[24] Cong Ma, Zhenjiang Miao, Min Li, Shaoyue Song, and Ming Hsuan Yang. Detecting anomalous trajectories via recurrent neural networks. In Greg Mori, Hongdong Li, C.V. Jawahar, and Konrad Schindler, editors, *Computer Vision – ACCV 2018 - 14th Asian Conference on Computer Vision, Revised Selected Papers*, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), pages 370–382, Germany, 2019. Springer Verlag. 14th Asian Conference on Computer Vision, ACCV 2018 ; Conference date: 02-12-2018 Through 06-12-2018.

[25] Neelu Madan, Kamal Nasrollahi, and Thomas B. Moeslund. Attention-enabled object detection to improve one-stage tracker. In *Accepted in Intelligent Systems Conference (IntelliSys)*, 2021.

[26] Michael Mathieu, Camille Couprie, and Yann LeCun. Deep multi-scale video prediction beyond mean square error. Jan. 2016. 4th International Conference on Learning Representations, ICLR 2016 ; Conference date: 02-05-2016 Through 04-05-2016.

[27] Romero Morais, Vuong Le, Truyen Tran, Budhaditya Saha, Moussa Mansour, and Svetha Venkatesh. Learning regularity in skeleton trajectories for anomaly detection in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[28] Naohiko Suzuki, Kosuke Hirasawa, Kenichi Tanaka, Yoshinori Kobayashi, Yoichi Sato, and Yozo Fujino. Learning motion patterns and anomaly detection by human trajectory analysis. In *2007 IEEE International Conference on Systems, Man and Cybernetics*, pages 498–503, 2007.

[29] Trong-Nguyen Nguyen and Jean Meunier. Anomaly detection in video sequence with appearance-motion correspondence. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1273–1283, 2019.

[30] Hyunjong Park, Jongyoun Noh, and Bumsub Ham. Learning memory-guided normality for anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14372–14381, 2020.

[31] B. Ramachandra, Michael J. Jones, and Ranga Raju Vatsavai. A survey of single-scene video anomaly detection. *IEEE transactions on pattern analysis and machine intelligence*, PP, 2020.

[32] Mahdyar Ravanbakhsh, Moin Nabi, Enver Sangineto, Lucio Marcenaro, Carlo Regazzoni, and Nicu Sebe. Abnormal event detection in videos using generative adversarial nets. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 1577–1581. IEEE, 2017.

[33] Royston Rodrigues, Neha Bhargava, Rajbabu Velmurugan, and Subhasis Chaudhuri. Multi-timescale trajectory prediction for abnormal human activity detection. In *The IEEE Winter Conference on Applications of Computer Vision (WACV)*, March 2020.

[34] Yao Tang, Lin Zhao, Shanshan Zhang, Chen Gong, Guangyu Li, and Jian Yang. Integrating prediction and reconstruction for anomaly detection. *Pattern Recognition Letters*, 129:123–130, 2020.

[35] D. Yao, C. Zhang, Z. Zhu, J. Huang, and J. Bi. Trajectory clustering via deep representation learning. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 3880–3887, 2017.

[36] Guang Yu, Siqi Wang, Zhiping Cai, En Zhu, Chuanfu Xu, Jianping Yin, and Marius Kloft. Cloze test helps: Effective video anomaly detection via learning to complete video events. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 583–591, 2020.

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