Anomaly Detection in Videos Recorded by Drones in a Surveillance Context

Jordan Henrio and Tomoharu Nakashima
Osaka Prefecture University, Japan
Telephone: +81-72-254-9351
jordan.henrio@cs.osakafu-u.ac.jp
tomoharu.nakashima@kis.osakafu-u.ac.jp

Abstract—**Automatic anomaly detection is an important difficulty in various domains such as surveillance, medicine and industry. Current surveillance systems generally use fixed-position cameras that cannot track anomalies. By virtue of recent industrial developments, drones are increasingly employed in various domains. They can also be employed for surveillance as moving cameras. Nevertheless, the recent literature lacks contributions describing anomaly detection in videos recorded by drones or for those with dynamic backgrounds. This paper presents specific examination of anomaly detection on the mini-drone video dataset which consists of surveillance videos recorded by a drone. The suggested anomaly detector is a deep neural network composed of a convolutional neural network and a recurrent neural network, trained using supervised learning. Experiments were conducted on both the mini-drone video dataset and the more popular UMN dataset. Although our model achieves state-of-the-art results on the UMN dataset, it is less data-efficient. However, the main goal of this study is to promote research efforts to resolve the difficulty of anomaly detection in videos with dynamic background and to provide first baseline results for additional contributions on the mini-drone video dataset.**

1. Introduction

From a general perspective, an anomaly is a pattern that differs from a standard pattern. Therefore, anomalies depend on the phenomenon of interest. For instance, in the surveillance case, an abnormal event might be an intrusion. In medicine, it might be the presence of a tumor. Neglected anomalies can present costly ramifications. Therefore, their detection is an important task that arises in various domains. However, by definition, an anomaly is a rare event. For this reason, their detection is tedious. For example, in the case of surveillance, human operators must watch continuous video streams carefully, possibly from several sources, for hours. For this reason, an automatic system that performs this task is desired.

Current surveillance systems commonly employ fixed-position cameras that present some difficulty in anomaly tracking. By virtue of recent industrial developments, drones are increasingly used in various applications related to agriculture, traffic monitoring, construction industry, and disaster area investigations. In the context of surveillance, drones are useful as mobile cameras. They overcome weaknesses of fixed-position cameras. It is particularly interesting that no recent reports describe studies addressing the difficulty of finding anomalies in videos with a dynamic background such as those recorded by drones. Recent contributions in anomaly detection for the surveillance case rely upon the assumption of a static background using data representations such as optical flows [1]. However, such methods are difficult to apply for videos with dynamic backgrounds.

For decades, information extraction from videos has been done using handcrafted features. However, recent advances in deep learning and the huge amount of publicly available data have made it possible to build systems that learn to extract useful features automatically. Consequently, tools such as convolutional neural networks (CNN) and recurrent neural networks (RNN) are promising for extraction of spatiotemporal information from videos. As described herein, we propose to train, with a supervised learning approach, a model that combines both a CNN and an RNN for the task of anomaly detection.

Numerous publicly available datasets have been designed for anomaly detection in the context of surveillance. However, few of those include videos recorded by drones. As described in this paper, we specifically examine the mini-drone video (MDV) dataset [2]. This report is the first of the relevant literature describing a contribution for anomaly detection using this dataset. Additionally, we evaluate our method on Detection of Unusual Crowd Activity dataset provided by the University of Minnesota (UMN) [3] which is more commonly used in the literature.

The difference between the results obtained on the MDV dataset and those obtained on the UMN dataset indicates the former as more complex, both in varieties of scenes and conditions under which videos were recorded. Furthermore, the model struggles to detect anomalies that are similar to normal patterns. However, although our method is less data-efficient than unsupervised learning approaches, we obtained state-of-the-art results on the UMN dataset.

Our contribution is to deal with anomaly detection on a more complex dataset than those commonly investigated. As a first trial, we propose a simple model combining a CNN and an RNN. The first results suggest there are rooms of improvement. Therefore, we hope our results stand as a

2577-1655/18/$31.00 ©2018 IEEE
DOI 10.1109/SMC.2018.00429

baseline for additional studies of this dataset.

This paper is organized as follows. Section 2 presents a review of recent contributions on anomaly detection. Section 3 presents a description of the proposed method. Section 4 explains the MDV dataset. The model is evaluated on both MDV and UMN datasets in section 5. Then, we conclude the discussion in this paper and suggest possible methods of improvement for additional studies.

2. Related Work

By definition, anomalies are rare events. Therefore, a natural idea is to estimate the data distribution and to detect outliers. Sun et al. [4] proposed an online growing neural gas that adapts itself to changing environments to estimate the data distribution topology. During tests, when the neuron representing a given input video frame is far away the densest parts of the gas, the corresponding frame is regarded as abnormal. Another approach was suggested by Xu et al. [1], who proposed the use of three denoising auto-encoders to extract appearance, motion and aggregation of both features from video frames. Then, three one-class Support Vector Machines (SVM) are trained for anomaly detection with features extracted using auto-encoders. During tests, the scores of the three SVMs are aggregated to compute the global abnormal score of a given video frame.

Another common approach presented in the literature is to learn the representation basis of normal patterns, and then to try to reconstruct an input frame based on this representation. In this case, a frame providing a large reconstruction error is likely to be abnormal. Ren et al. [5] proposed to learn dictionaries representing normal patterns. Chong et al. [6] reported the use of a spatiotemporal auto-encoder with an encoder consisting of convolutional layers and a decoder of deconvolutional layers. To extract temporal features, the auto-encoder's bottleneck is made with convolutional long short-term memory (conv-LSTM) units. Hasan et al. [7] proposed a similar approach, except that the model includes no conv-LSTM units. Rather, to extract temporal information, the method inputs a stack of several consecutive frames to the auto-encoder so that the input volume has a temporal dimension. The idea that large reconstruction errors are useful for anomaly detection has also been investigated in a medical context by Schlegl et al. [8] for retinal fluid or hyper-reflective foci detection. To do so, the authors suggest the use of a Generative Adversarial Network (GAN). Because GAN only learns to map latent representations to images, the authors also suggest learning reverse mapping, so that a given image can be reconstructed and evaluated during testing.

Recently, Munawar et al. [9] defined anomalies as observed events that differ from expectations. Consequently, they suggested comparison of the prediction of a model and the actual observation. A model is trained to predict a representation of the frame at time-step $t + 1$ given the frame at time-step t.

In this work, we specifically examines the mini-drone video dataset for which the training set includes both normal and abnormal patterns. Therefore, we decided to work first on training a model under a supervised learning approach to propose baseline results. All methods cited above tackle the anomaly detection problem with semi-supervised or unsupervised learning approaches. Concretely, they use only normal patterns during training and try to detect outliers during tests. Manuwar et al. [10] investigated supervised learning for anomaly detection based on reconstruction error. They suggested to train a model by alternating positive and negative learning phases. A positive learning phase searches for the parameters that minimizes the reconstruction error of normal frames, while a negative phase maximizes the reconstruction error of abnormal frames.

Mini-drone video dataset's anomalies are essentially determined by actors' actions. Thus, one can consider the problem as action detection. Hong et al. suggested methods to extract the skeleton from the silhouette of an actor by using sparse coding in [11], and a multimodal deep auto-encoder in [12]. However, the mini-drone video dataset includes a few examples per action. Furthermore, some actions are present in the test set but not in the train set. Thus, action detection is hardly conceivable. Consequently, this work does not attempt to recognize each action, but to classify them as normal or not.

3. Proposed Method

3.1. Convolutional Neural Network

The first component of our model is designed to extract visual (spatial) information contained in video frames. To do so, we use VGG-16, a famous CNN architecture proposed by Simonyan and Zisserman [13]. This architecture includes thirteen convolutional layers, five max pooling operations and three fully connected layers as depicted in Figure 1.

Training a deep CNN such as VGG-16 requires a large dataset and requires much time. Consequently, the proposed method uses transfer learning to alleviate the training process by initializing the weights of VGG-16 to those resulting from training on the ImageNet [14] dataset. This dataset includes more than fourteen million images divided up into one thousand classes, making it a challenging natural image classification dataset.

3.2. Recurrent Neural Network

The second component of our model is aimed at extracting temporal information contained between video frames. Standard neural networks and CNN do not extract time dependencies between the current and previous input features. A common approach for the extraction of temporal information is to employ an RNN. This class of neural network extracts information from the current input while considering features from earlier inputs.

Specifically, the proposed method employs a Long Short-Term Memory (LSTM) [15]. This model is commonly used to cope with the vanishing/exploding gradient problem

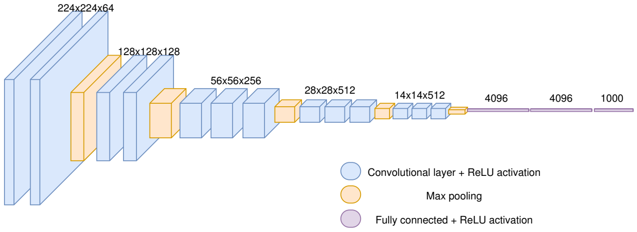

Figure 1. VGG16 architecture (best viewed in color).

commonly encountered with standard RNN. LSTM tackles this problem, by using gates that maintain the recurrent state c by removing or adding information according to the current input v_t and the previous output h_{t-1}. Figure 2 provides a schematic view of this process.

3.3. Proposed Model

In this work, we specifically examine the difficulty of anomaly detection as a binary classification task. The actual model proposed in this paper is the combination of the two components presented in earlier sections. Additionally, it ends up with a final fully connected layer that computes evidence for the classes ($normal = 0$ or $abnormal = 1$). Here, it consists of a single neuron with a sigmoid activation. Figure 3 depicts an abstract view of the proposed model.

It is noteworthy that the VGG-16 actually employed here is slightly different from the one represented in Figure 1. The last fully connected layer of the original VGG16 comprises 1000 neurons that compute evidence for the classes of the ImageNet dataset. We need no such evidence for our objective. Therefore, this final layer was simply removed.

In this work, we examine the training of the proposed neural network with a supervised learning approach. It is trained according to the Binary Cross Entropy loss function l defined in (1) as

$$l(y, \hat{y}) = -\frac{1}{n}\sum_{i=1}^{n}[y_i \log(\hat{y}_i) + (1 - y_i)(\log(1 - \hat{y}_i))], \quad (1)$$

where y is a vector containing the ground-truth labels, $\hat{y}$ is a vector containing the model predictions, and n represents the number of training examples.

The model is trained using the Adam optimizer [16] and the early stopping algorithm presented in [17], which interrupts the training process automatically when the loss on the test set does not decrease for 10 consecutive epochs.

To help the model to generalize from the training data, dropout [18] is used in the LSTM. Furthermore, the training set was augmented online by applying random transformations such as random cropping, left-right flip and pixel dropout.

To match the VGG-16 architecture, input frames are resized to 224×224. Additionally, they are individually zero-centered and standardized.

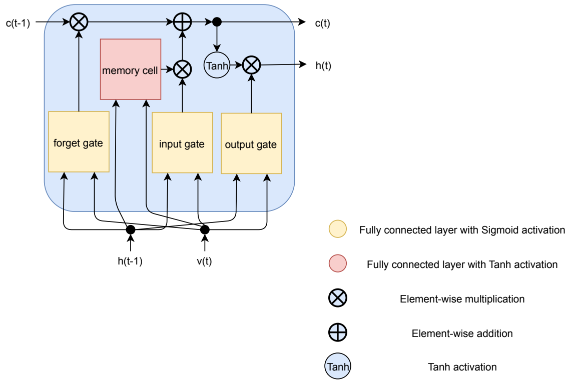
c(t-1)
c(t)
memory cell
Tanh
h(t)
forget gate
input gate
output gate
h(t-1)
v(t)
Fully connected layer with Sigmoid activation
Fully connected layer with Tanh activation
Element-wise multiplication
Element-wise addition
Tanh
Tanh activation

Figure 2. LSTM diagram (best viewed in color).

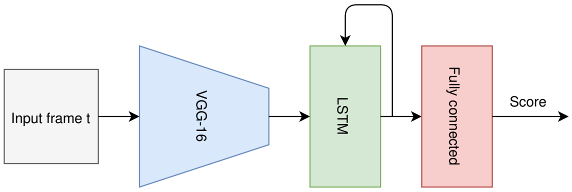

Figure 3. Diagram representing the suggested model.

4. Mini-drone Video dataset

On the one hand, the mini-drone video (MDV) dataset [2] was originally proposed for the design of privacy filtering methods. On the other hand, scenes that constitute it have three types: *normal*, *suspicious* and *abnormal*. Consequently, it could be used in the context of anomaly detection. Furthermore, it has been annotated with the ViPER-GT tool [19]. Therefore, each dataset's video is associated with an XML file which describes the different elements present in that video and their position in each frame. Consequently, the dataset can also be used for object detection and object tracking tasks.

The dataset contains 38 videos recorded by a drone flying at a low altitude. The video duration is 16–24 s. In addition, the set is divided into two subsets: $train$ and $test$. The training set includes 15 videos (9,497 frames), whereas the test set contains 23 videos (13,798 frames).

Scenes take place in a parking lot and represent various possible situations. The dataset proposes three situation categories: *normal*, *suspicious* and *abnormal*. Categories are almost all determined by the actions of the persons involved in the videos. Normal scenes represent various situations such as persons walking, conversing or parking their car. Abnormal scenes represent situations that are either illicit, such as fighting, or at least dangerous such as a person falling down. Suspicious scenes represent situations that contain non-illicit actions, but which would draw the surveillance staff's attention. For instance, a person is wandering in the parking lot and is looking into cars through the windows. Such a situation suggests that the person is looking for a car to steal. In addition, the dataset also contains anomalous videos that do not stage any human action. In such videos, the anomalies are produced by objects as cars parked outside a parking spot. However, in this work, we ignore such anomalies and treat them as normal patterns. Table 1 lists all the different actions present in the dataset and how we categorized them.

It is noteworthy that the label *repairing* is not present in the original dataset. The videos named *Broken_CloseUp_Day_Half_1_1_1* and *Broken_CloseUp_Day_Half_1_1_2* stage a person repairing his car, but the human action was not labeled in the corresponding XML files. Therefore, we manually labeled the empty human actions in the XML files corresponding to these two videos.

In addition, videos were recorded with different movements and positions of the drone. The altitude can vary between videos. In some of them, the drone approaches a specific target. In some videos the drone remains motionless, but it follows a particular target in others. The dataset also includes videos in which the drone circles a parking slot to obtain global information. Furthermore, the videos were recorded at different times of day: even at night. For that reason, differences in luminosity exist among videos.

Consequently, the MDV dataset is complex both in terms of the variety of scenes it includes and the different conditions under which the videos were recorded.

5. Experiments

5.1. MDV dataset baseline results

As a first experiment, the proposed model was evaluated on the MDV dataset to provide baseline results for future studies. To do so, we specifically examined two scenarios. In the first case, we considered suspicious scenes as normal events (*suspicious* = 0), although suspicious scenes were regarded as abnormal (*suspicious* = 1) in the second case.

For this experiment, we used a pre-trained VGG-16 without fine-tuning it and an LSTM with one layer of 256 neurons. The model was trained with a learning rate of 0.00001 and a dropout rate of 0.5.

The resulting models were evaluated quantitatively according to the Area Under the Receiver-Operator Curve (AUC) score on the test set. As presented in Table 2, the proposed method (VGG16 + LSTM) obtained an AUC score of 0.7275 in the first case, although the score was lower in the second scenario: 0.6445. The gap separating the AUC scores indicates that the proposed model struggles to distinguish suspicious human behaviors from normal ones.

In addition, the model failed to detect abnormal scenes that are not present in the training set, such as a man attacking another man to steal his car, or a collision involving a pedestrian and a cyclist. This failure highlights the limitations of supervised learning for anomaly detection difficulties for which examples of abnormal patterns are rare.

As a comparison point, we trained a convolutional autoencoder by alternating positive and negative learning phases as proposed in [10]. We implemented an encoder whose architecture is the same as VGG16 (without fully connected layer) and the decoder is the symmetric architecture with max-pooling operations replaced with bilinear upsampling operators. This model obtained an AUC score of 0.6209 when considering suspicious human behaviors as normal and 0.6232 when considering them as abnormal. While the performance are lower than those of VGG16 + LSTM, it is noteworthy that the auto-encoder did not use any temporal features.

TABLE 1. List of the Actions in the Mini-drone Video dataset

Category	Action
normal	walking, standing, talking, nothing, parking, parked, moving, stopping
suspicious	loitering
abnormal	fighting, picking up, attacking, stealing, cycling, running, falling, repairing

5.2. MDV dataset ablation study

To verify the importance of temporal features on the MDV dataset, we removed the LSTM from our model. Thus, in this setting, the model has to detect anomalies only from independent snapshots. As expected, we observed a considerable loss on AUC score. We obtained an AUC score of 0.5012 when considering suspicious behaviors as normal and 0.4999 when considering suspicious behaviors as abnormal. Thus, without temporal information the proposed model performs not better than a random classifier.

When using temporal features (VGG16 + LSTM), we observed that the model is sensitive to the drone movements, especially when the drone chases a particular target. For example, video *Suspicious_Follow_Day_Half_0_f_1* portrays a person walking on the parking lot and looking at the window of a car, perhaps to steal it. Around frame 250, he realizes the presence of the drone and starts to walk away. Then, the drone chases him. As depicted on Figure 4 the model's score starts to increase. Consequently, the dynamic background property resulting from drone's movement acts as cue for the presence of an abnormal event. The dashed line in Figure 4 represents the threshold that maximizes classification accuracy.

5.3. UMN dataset

As a second set of experiments, we evaluated our method on the Detection of Unusual Crowd Activity dataset provided by the University of Minnesota (UMN) [3] to compare it with other contributions of the domain because this dataset has been studied widely by the anomaly detection community. However, we do not evaluate our method on other popular datasets such as UCSD [20] or Avenue [21] because their training set contains only normal patterns, rendering our supervised learning approach unusable without considerably changing their distribution.

TABLE 2. AUC SCORE COMPARISONS FOR THE MDV DATASET

Model	Set	AUC score
VGG16 + LSTM	*suspicious* = 0	**0.7275**
	suspicious = 1	**0.6445**
Negative Learning [10]	*suspicious* = 0	0.6209
	suspicious = 1	0.6232
VGG16	*suspicious* = 0	0.5012
	suspicious = 1	0.4999

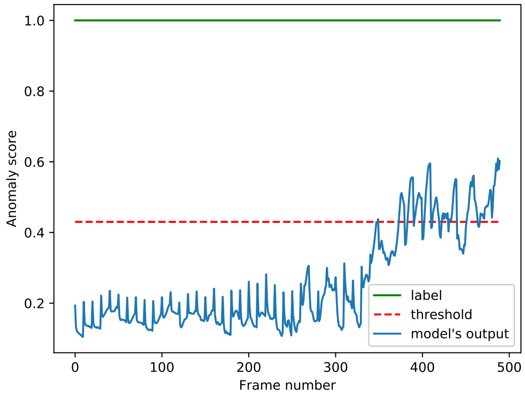

Figure 4. Output score's sensitivity according to the drone's movement.

The UMN dataset consists of a single video formed by 15,478 frames. The video is divisible into several scenes that take place in three stages. Each scene represents people walking and then suddenly running away from something. The dataset is not divided properly into a training and test set. Since we train our model using a supervised learning approach, it is necessary to include both examples of normal and abnormal patterns in the training set. Additionally, we need examples for the three stages present in the video. Therefore, we build our training set using the whole first scene of each stage (frames [0, 624], [1453, 2001] and [5596, 6253]). The test set comprises all remaining frames. The resulting training set includes 1832 frames. The test set includes 5907 frames.

From this experiment, we ascertained that a pre-trained VGG-16 without fine-tuning and an LSTM with one layer of 16 neurons was sufficient. The model was trained with a learning rate of 0.0001 and a dropout rate of 0.5.

After training the model during 200 epochs, we found the threshold 0.8547 to maximize accuracy on the test set (0.9954). In addition, this model caused an AUC score of 0.9994, which is similar to other state-of-the-art contributions on this dataset. Table 3 presents a summary the AUC score of recent contributions on the UMN dataset. Our results are similar to those of Sun et al. [4], but our method is unfair. All the contributions present in Table 3 employ unsupervised learning approaches and train only on normal patterns. However, in our case, we include both normal and abnormal patterns in our training set. Consequently, our method is less data-efficient.

The difference between the results obtained on the MDV dataset and those obtained on the UMN dataset suggest that the latter is too simple. First, it does not contain widely various scenes. Furthermore, anomalies can be distinguished easily from normal patterns because the whole crowd's behavior changes. Secondly, anomalies are indicated directly in the frames by a visual marker, as depicted in Figure 5. This marker generally appears several frames after the crowd starts running. Consequently, if this marker were used to label the frames, then one could consider the resulting labels as inaccurate. We also labeled the frames by hand without considering the visual marker. In doing so, the AUC score of our model decreased to 0.92. When using the visual marker as a label, our model seems to ignore the crowd behavior and only bases its prediction on the presence of the visual marker or not. This fact is represented in Figure 6, where the signal *output* fits the signal *marker labels*, although a human operator can detect the anomaly as represented by the signal *hand-made labels*. Small pikes between frames 300 and 500 seem to correspond to a change of luminosity.

6. Conclusion

As described herein, we specifically examined anomaly detection in a surveillance context, especially for the mini-drone video dataset that consists of surveillance videos recorded by a drone.

The model proposed in this paper was evaluated on the MDV dataset. This report is the first of an attempt to perform anomaly detection with this dataset. Additionally, the model was evaluated with the UMN dataset, which is more commonly used within the anomaly detection community. Despite its simplicity, the model reached state-of-the-art performance on the latter dataset. The difference of performances observed on the MDV and the UMN datasets suggests that the former is more complex in terms of situation variety.

Furthermore, we observed the model to be sensitive to the dynamic background of the MDV dataset's videos. Consequently, our future work will specifically examine the development of techniques that reduce this effect.

Finally, our results suggest that supervised learning is not well suited for anomaly detection. Although the model performs well on the UMN dataset, other methods yield similar results obtained using no anomaly example. Furthermore, the model failed to detect situations it never encountered during training on the MDV dataset, which is problematic because it is difficult, if not impossible, to compile a dataset representing all possible anomalies by many instances. Consequently, our future work will specifically examine the design of models trained by unsupervised learning to alleviate the need for labeled anomalies.

Acknowledgments

This research was partly supported by the ICOM Foundation, Japan.

TABLE 3. AUC SCORE COMPARISONS ON THE UMN DATASET

Method	AUC score
Sun et al. [4]	0.9965
Ravanbakhsh et al. [22]	0.99
Ionescu et al. [23]	0.951
Del Giorno et al. [24]	0.91
Proposed method	**0.9994**

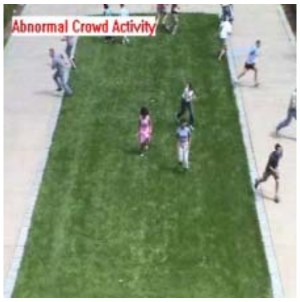

Figure 5. Example of UMN's anomaly visual marker.

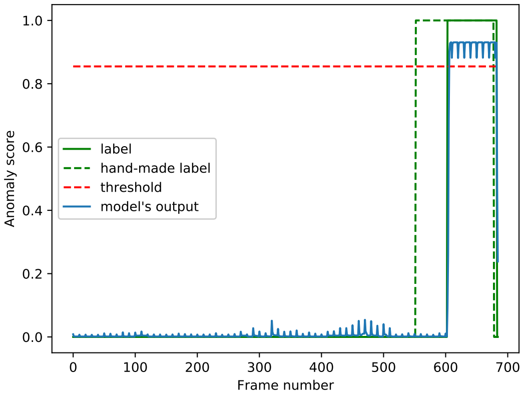

Figure 6. Comparison of visual markers and hand-made labels.

References

[1] Dan Xu, Yan Yan, Elisa Ricci, and Nicu Sebe. Detecting anomalous events in videos by learning deep representations of appearance and motion. *Computer Vision and Image Understanding*, 156(Supplement C):117–127, 2017. Image and Video Understanding in Big Data.

[2] Margherita Bonetto, Pavel Korshunov, Giovanni Ramponi, and Touradj Ebrahimi. Privacy in mini-drone based video surveillance. *2015 IEEE International Conference on Image Processing (ICIP)*, pages 2464–2469, 2015.

[3] Department of Computer Science University of Minnesota and Engineering. Detection of unusual crowd activity. Video dataset. http://mha.cs.umn.edu/Movies/Crowd-Activity-All.avi.

[4] Qianru Sun, Hong Liu, and Tatsuya Harada. Online growing neural gas for anomaly detection in changing surveillance scenes. *Pattern Recognition*, 64(Supplement C):187–201, 2017.

[5] Huamin Ren, Weifeng Liu, Søren Ingvor Olsen, Sergio Escalera, and Thomas B Moeslund. Unsupervised behavior-specific dictionary learning for abnormal event detection. In *BMVC*, pages 28.1–28.13, 2015.

[6] Yong Shean Chong and Yong Haur Tay. *Abnormal Event Detection in Videos Using Spatiotemporal Autoencoder*, pages 189–196. Springer International Publishing, Cham, 2017.

[7] 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*, pages 733–742, 2016.

[8] Thomas Schlegl, Philipp Seeböck, Sebastian M. Waldstein, Ursula Schmidt-Erfurth, and Georg Langs. *Unsupervised Anomaly Detection with Generative Adversarial Networks to Guide Marker Discovery*, pages 146–157. Springer International Publishing, Cham, 2017.

[9] Asim Munawar, Phongtharin Vinayavekhin, and Giovanni De Magistris. Spatiotemporal anomaly detection for industrial robots through prediction in unsupervised feature space. In *Applications of Computer Vision (WACV), 2017 IEEE Winter Conference on*, pages 1017–1025, 2017.

[10] Asim Munawar, Phongtharin Vinayavekhin, and Giovanni De Magistris. Limiting the reconstruction capability of generative neural network using negative learning. *2017 IEEE 27th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6, 2017.

[11] C. Hong, J. Yu, D. Tao, and M. Wang. Image-based three-dimensional human pose recovery by multiview locality-sensitive sparse retrieval. *IEEE Transactions on Industrial Electronics*, 62(6):3742–3751, June 2015.

[12] C. Hong, J. Yu, J. Wan, D. Tao, and M. Wang. Multimodal deep autoencoder for human pose recovery. *IEEE Transactions on Image Processing*, 24(12):5659–5670, December 2015.

[13] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[14] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

[15] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, November 1997.

[16] Diederik P. Kingma and Jimmy Lei Ba. Adam: A method for stochastic optimization. In *The Third International Conference for Learning Representations*, 2015.

[17] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*, chapter 7, pages 241–249. MIT Press, 2016.

[18] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15:1929–1958, 2014.

[19] D. Doermann and D. Mihalcik. Tools and techniques for video performance evaluation. In *Proceedings 15th International Conference on Pattern Recognition. ICPR-2000*, pages 167–170, 2000.

[20] Statistical Visual Computing Laboratory University of California San Diego. Ucsd anomaly detection dataset. Video dataset. http://www.svcl.ucsd.edu/projects/anomaly/dataset.htm.

[21] Cewu Lu, Jianping Shi, and Jiaya Jia. Abnormal event detection at 150 fps in matlab. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2720–2727, 2013.

[22] Mahdyar Ravanbakhsh, Enver Sangineto, Moin Nabi, and Nicu Sebe. Training adversarial discriminators for cross-channel abnormal event detection in crowds. *CoRR*, abs/1706.07680, 2017.

[23] Radu Tudor Ionescu, Sorina Smeureanu, Bogdan Alexe, and Marius Popescu. Unmasking the abnormal events in video. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2914–2922, 2017.

[24] Allison Del Giorno, J Andrew Bagnell, and Martial Hebert. A discriminative framework for anomaly detection in large videos. In *European Conference on Computer Vision*, pages 334–349. Springer, 2016.