Detection of Video Anomalies Using Convolutional Autoencoders and One-Class Support Vector Machines

Matheus Gutoski[1], Nelson Marcelo Romero Aquino[2]
Manassés Ribeiro[3], André Engênio Lazzaretti[4], Heitor Silvério Lopes[5]

Federal University of Technology - Paraná
Av. Sete de Setembro, 3165 - Rebouças - Curitiba, 80230-901, Brazil
Catarinense Federal Institute of Education, Science and Technology –(IFC),
Rod. SC 135 km 125, Videira (SC) 89560-000, Brazil
[1]matheusgutoski@gmail.com
[2]nmarceloromero@gmail.com
[3]manasses@ifc-videira.edu.br
[4]lazzaretti@utfpr.edu.br
[5]hslopes@utfpr.edu.br

Abstract. With the growth of image data being generated by surveillance cameras, automated video analysis has become necessary in order to detect unusual events. Recently, Deep Learning methods have achieved the state of the art results in many tasks related to computer vision. Among Deep Learning methods, the Autoencoder is commonly used for anomaly detection tasks. This work presents a method to classify frames of four different well known video datasets as normal or anomalous by using reconstruction errors as features for a classifier. To perform this task, Convolutional Autoencoders and One-Class SVMs were employed. Results suggest that the method is capable of detecting anomalies across the four different benchmark datasets. We also present a comparison with the state of the art approaches and data visualization.

1 INTRODUCTION

In recent years, the amount of data generated by surveillance systems has grown due to the decreasing cost of image capturing devices and the elevated concern with security [14]. However, the volume of data grew much faster than the availability of human observers, thus leading to a serious problem. A possible solution to this problem is the development of automated video surveillance systems.

Anomaly detection, also known as outlier detection, is a well known problem within the Pattern Recognition field [14]. It can be defined as a recognition problem, such as the pattern to be recognized is is scarce or not present in the training data [9].

In video anomaly detection, an abnormality frequently reported is the unexpected crowd behavior. The most common approach to detect abnormality in videos is by training a classifier using data containing only normal situations, which is then used to detect deviations [7].

Several models have been proposed to address the video anomaly detection problem [6, 2, 13]. Many of the proposed models use handcrafted feature extractors, which are usually specific to a given application and may not achieve satisfactory performance for other applications. Deep Learning models, on the other hand, have the advantage of learning suitable features, as well as the classifier at the same time. Recent Deep Learning based models have achieved good performance on benchmark datasets [10, 16, 3]. These models mainly use Autoencoders and its variants, such as Stacked, Denoising and Convolutional Autoencoders.

Autoencoders are considered an unsupervised learning method, since their learning process do not require class labels. Their goal is to reconstruct the input data after going through one or more layers of decreasing complexity. The difference between the input and the output is known as the reconstruction error.

The reconstruction error carries information that can be used to discriminate between normal and abnormal frames, as suggested by [10]. Smaller reconstruction errors are expected for normal instances, since they happen more often in the training set. Therefore, frames with high reconstruction error are assumed to be abnormal.

To test this hypothesis, this work presents a Convolutional Autoencoder video anomaly detector. The goal is to determine whether frames of a video sequence are normal or anomalous by using the reconstruction error. The classification is performed by One-Class SVMs. The models are trained, tested and evaluated using four well known benchmark datasets. This work is an extension of the study presented in Ribeiro et al. [10], since it also uses the strategy of fusing low-level features with high level appearance and motion features. The main contribution of this work is the use of the reconstruction errors obtained by the fusion strategy as features for a One-Class SVM classifier. This approach also provides a three-dimensional visualization of the data.

This work is organized as follows: Section 2 presents some theoretical aspects of this work. Section 3 describes the problem and the methods employed, from feature extraction to evaluation. Section 4 presents the experiments. Section 5 shows the results obtained using the proposed methodology. Section 6 presents a brief discussion about the difficulties and the results achieved.

2 Theoretical Aspects

2.1 Auto-encoders

Introduced by Rumelhart et al. [11], the Autoencoder is a fully connected neural network with one hidden layer. The main aspect of this network is the learning from unlabeled data. This is done by reconstructing the input at the output layer. The Autoencoder receives an input $\mathbf{x} \in \mathbb{R}^d$ and maps it to the hidden layer $\mathbf{h} \in \mathbb{R}^{d'}$ with the mapping function $\mathbf{h} = f_\Theta = \sigma(\mathbf{W}\mathbf{x} + b)$ using parameters

$\Theta = \{\mathbf{W}, b\}$. The first part of the network defines the encoder, and the second part defines the decoder. The decoder parameters are originally the transposed parameters of the encoder $\mathbf{W'} = \mathbf{W}^T$ [8].

Since the Autoencoder is an unsupervised learning method, it does not require class labels. It minimizes the reconstruction error e between input $\mathbf{x}_i$ and output $\mathbf{y}_i$ by adjusting its parameters as shown by the loss function in Equation 1

$$e(\mathbf{x}, \mathbf{y}) = \frac{1}{2N} \sum_{i=1}^{N} \|\mathbf{x}_i - \mathbf{y}_i\|_2^2. \tag{1}$$

2.2 Convolutional auto-encoder

The Autoencoder is limited by its inability to capture the 2D image structure [8], which is an important aspect of video anomaly detection, since anomalies occur in specific locations in the scene. An alternative way to tackle this issue is to use the Convolutional Autoencoder (CAE), proposed by [8]. The CAE is able of capturing the 2D image structure since its weights are shared among all locations in the input image. Equation 2 shows the loss function used in the CAE.

$$e(\mathbf{x}, \mathbf{y}, \mathbf{W}) = \frac{1}{2N} \sum_{i=1}^{N} \|\mathbf{x}_i - \mathbf{y}_i\|_2^2 + \lambda \left\|\mathbf{W}\right\|_2^2, \tag{2}$$

where λ is the regularization parameter for the regularization term $\|\mathbf{W}\|_2^2$, normally used during the training procedure of the CAE.

3 METHODOLOGY

Video anomaly detection is often divided in two main categories: frame level detection and pixel level detection. In frame level detection, the goal is to classify a frame as being normal or abnormal. In pixel level detection, the objective is to find the anomaly somewhere in the frame, similarly to a segmentation task. This work will address the frame level detection category. Figure 1 shows the proposed methodology.

3.1 Preprocessing

To prepare the data, each dataset was treated separately, and video sequences were converted into individual frames. Each frame was resized to 235×155 pixels, to further apply the Canny Edge Detector [1], and calculate the Optical Flow [4] using the current and previous frame, to capture movement patterns.

The final data consists of three-dimensional arrays. The first dimension contains the gray scale image. The second dimension contains the same image frame filtered with the Canny algorithm. The third dimension contains the Optical Flow. Figure 2 displays a sample and each of the dimensions.

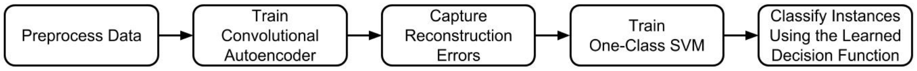

Preprocess Data
Train Convolutional Autoencoder
Capture Reconstruction Errors
Train One-Class SVM
Classify Instances Using the Learned Decision Function

Fig. 1. The proposed methodology

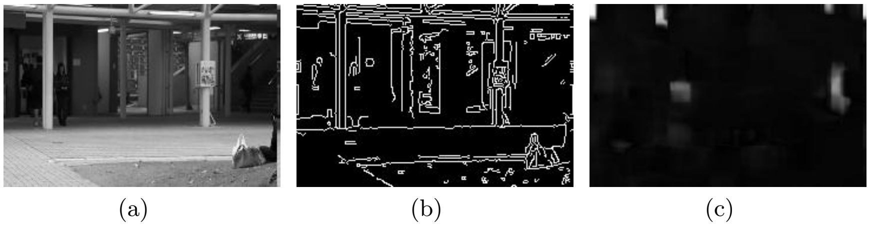


Fig. 2. Sample input from the Avenue dataset. Each figure represents a channel of the image fed to the CAE. Figure 2(a) shows the first channel, the gray scale image. Figure 2(b) shows the second channel, the edge filter of the image. Figure 2(c) shows the Optical Flow extracted from the current frame using the previous one.

3.2 Feature Extraction

The feature extraction process consists of two steps. The first step is training the CAE using the three-dimensional train set. The CAE is trained by minimizing the sum of the Euclidean distances between each pixel of the input and output images. CAEs are trained separately for each dataset. The train process stops after 1.000 epochs. The network architecture is based on that presented by Ribeiro et al. [10], since it has shown good performance, and its structure is as follows: the encoder part contains three convolutional and two max pooling layers, located after the first and the second convolutions. The decoder mirrors the encoder part. No tied weights were used to train the network. The hyperbolic tangent activation function is used after convolution and deconvolution layers. Training was done using the Adaptive Gradient Algorithm (AdaGrad) with a fixed learning rate of 0.0001 and L2 regularization. The network was implemented using the Caffe Deep Learning Framework [5] and trained using Graphics Processing Units (GPUs).

The second step is to forward both train and test sets throughout the trained network in order to capture the reconstruction errors. The errors are computed as the sum of the Euclidean distances of each pixel between the input data (original frame) and the output data (reconstructed frame). However, this time the reconstruction error is calculated on each channel individually.

The feature vector of each sample is composed of three attributes. The first attribute is the reconstruction error of the first channel (gray scale image). The second attribute is the reconstruction error of the second channel (edges detected with the Canny algorithm). The third attribute is the reconstruction error of the third channel (Optical Flow).

By using the appearance features (Canny), it is expected that anomalies that include new objects or shapes can be suitably represented. Furthermore, the Optical Flow may be able to capture anomalies related to movement.

3.3 Normalization

After computing the reconstruction errors, the final feature vector of both train and test sets are normalized by dividing each element by the Euclidean Norm of the train feature vector. The Euclidean norm is defined by $\|x\|_2 = \sqrt{x_1^2 + ... + x_n^2}$.

3.4 Classification Method

To perform classification, the One-Class SVM with the RBF kernel function was chosen. The classifier is trained with the three-feature train set in order to learn the normal patterns. Each sample of the test set is classified as normal or abnormal based on the decision function learned in the train phase. .

3.5 Evaluation

To evaluate the model, the area under the Receiver Operating Characteristic (ROC) curve and the Equal Error Rate (EER) were computed using the test set. The evaluation is performed using the test set. The state of the art of each dataset is also presented to provide means for comparison. Furthermore, confusion matrices and 3-D scatter plots are shown.

The ROC curve is calculated by using the distances from each point to the closest decision border found by the One-class SVM. Points within the decision borders have a positive sign, whereas points outside the decision borders have a negative sign. The default One-class SVM classification uses the threshold 0 to classify instances. By using different thresholds, it is possible to plot the ROC curve and calculate the EER. The EER is the point in the ROC curve with the best balance between true positives and false positives. However, the optimal threshold should be defined according to the needs of the user, since it is a trade-off between true positives and false positives.

4 EXPERIMENTS

The methodology proposed in Section 3 was applied to four different datasets. Each of these datasets provide their own train and test sets with ground truth information for evaluation.

The benchmark datasets used for the video anomaly detection problem in this work are: Avenue[1], UCSD Ped 1 and Ped 2[2], and UMN[3].

[1] http://www.cse.cuhk.edu.hk/leojia/projects/detectabnormal/dataset.html

[2] http://www.svcl.ucsd.edu/projects/anomaly/dataset.htm

[3] http://mha.cs.umn.edu/proj_events.shtml

4.1 Datasets

Avenue The Avenue dataset consists of 16 training videos (15,328 frames) and 21 testing videos (15,324 frames). The dataset contains a small amount of anomalies in the training set. Moreover, some normal situations rarely appear in the training videos. The test set contains both normal and anomalous events. Figure 3 displays anomalous events found within the test frames.

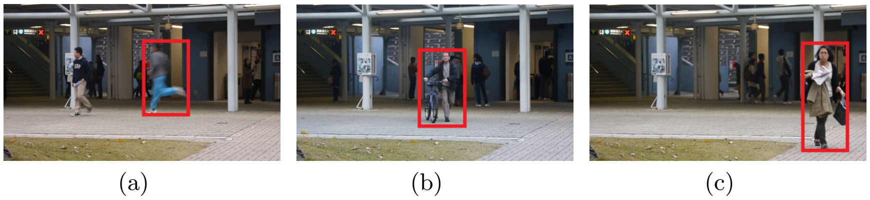


Fig. 3. Sample images from the Avenue dataset showing anomalous events. Figure 3(a) shows a strange action (running). Figure 3(b) shows an abnormal object in the scene (bike). Figure 3(c) shows a person walking in an unexpected direction.

UCSD The UCSD dataset is divided in two sub-datasets: ped 1 and ped 2.

The Ped 1 dataset contains 34 training videos (6,766 frames) and 10 labeled testing videos (1,990 frames). The videos contain pedestrians walking in a sidewalk, which is considered normal behavior. Anomalies are defined by small vehicles, bikes and skaters among pedestrians. Figure 4 shows an anomaly from the Ped 1 dataset.

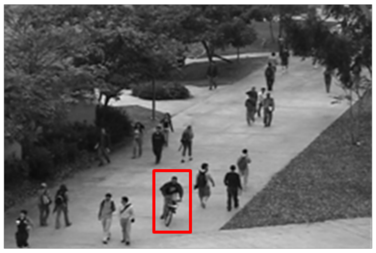
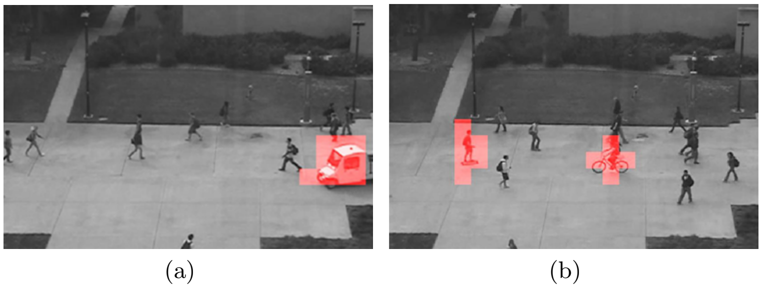


Fig. 4. Sample from the UCSD Ped 1 dataset. The red rectangle shows an anomaly (biker running amongst pedestrians)

The Ped 2 dataset contains 16 training videos (2,518 frames) and 12 testing videos (1,986 frames). Anomalies are the same as in ped 1. Figure 5 shows anomalous events in the ped 2 dataset.



(a) (b)

Fig. 5. Anomalies in the UCSD ped 2 dataset. Figure 5(a) shows a small vehicle amongst pedestrians. Figure 5(b) shows a biker and a skater.

UMN The UMN dataset consists of three different scenes, containing a total of 5,113 frames in the training set and 1,378 frames in the test set. In this dataset, anomalies are behavioral. Normal events are defined by people walking, and abnormal events occur when the crowd quickly evades the scene by running. Figure 6 shows a frame of each scene.

5 RESULTS

The results reported in this section regards the classification using the closest threshold to the ROC EER, which is a point of balance between true positives and false positives. Table 1 shows the classification results. True positive rates (TPR), False positive rates (FPR), Area under ROC (AUC) and the results of the State of the Art method Area under ROC (AUC SoA) are reported.

With the proposed method, the best results were obtained for the UMN dataset. Although the state of the art method for this dataset shows that the problem is not a very difficult one. However, achieving high values of AUC in a specific dataset requires a method fine tuned for a specific task, i.e, specifically designed for a certain dataset, whereas our method achieves similar performances across all datasets. So far, there is no approach capable to achieve the state-of-the-art results considering all the datasets

]

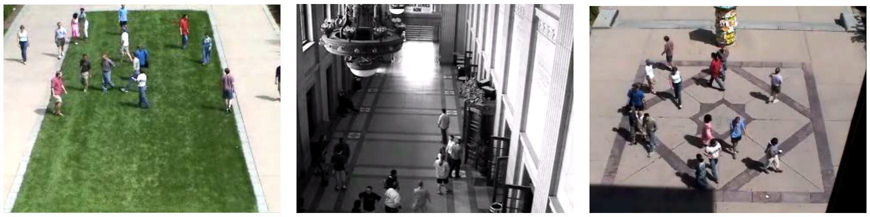


Fig. 6. Frames of each scene of the UMN dataset. These frames represent the normal crowd behavior.

Table 1. Classification results of four datasets. The classification results were obtained using the closest threshold to the Equal Error Rate (EER).

	Proposed Method			**State of the Art**	
Dataset	**TPR**	**FPR**	**AUC**	**AUC**	**Reference**
Avenue	0.67	0.38	0.69	0.77	[10]
UCSD Ped 1	0.53	0.37	0.59	0.92	[17]
UCSD Ped 2	0.81	0.60	0.61	0.908	[12]
UMN	0.79	0.33	0.81	0.99	[15]

The confusion matrices of each dataset are presented in Tables 2, 3, 4, 5, and the classification was done using the EER.

Table 2. Confusion Matrix of the Avenue Dataset at EER threshold

		Predicted	
		Normal	Anomaly
True	Normal	7848	3746
	Anomaly	1421	2288

Table 3. Confusion Matrix of the UCSD Ped 1 Dataset at EER threshold

		Predicted	
		Normal	Anomaly
True	Normal	401	356
	Anomaly	458	775

Table 4. Confusion Matrix of the UCSD Ped 2 Dataset at EER threshold

		Predicted	
		Normal	Anomaly
True	Normal	291	66
	Anomaly	993	648

Table 5. Confusion Matrix of the UMN Dataset at EER threshold

		Predicted	
		Normal	Anomaly
True	Normal	748	194
	Anomaly	147	289

One of the advantages of the proposed method is that it allows the visualization of the feature space through a three-dimensional scatter plot, since only three features are extracted from the videos (each feature is the reconstruction error of one of the three channels of the input image). By analyzing the feature space, some interesting aspects of the datasets can be observed. For instance, the UCSD Ped 1 dataset, as shown in Figure 7 (bottom); has high variance on the edge axis (Canny Edge Detector), which indicates a wide range of different objects in the scene, such as vehicles, bikes and pedestrians. The Avenue dataset contains most variations in the optical flow axis, as shown in Figure 7 (top). This indicates that anomalies in this dataset could be mostly related to movement.

In the UCSD Ped2 dataset, shown in Figure 8 (top), it can be observed a high variance in all three axis, indicating that high reconstruction errors are caused not only by movement patterns, but also by objects. Other interesting

Fig. 7. Each row of the image represents a different dataset. Top : Avenue, Bottom: UCSD Ped1. Each column represents: (left) Test set feature space (Ground Truth classification), (right) Test set feature space (One-class SVM EER prediction). Normal frames are represented by green dots, whilst anomalous frames are represented by red dots.

aspects of the datasets can be noticed. For instance, the UMN dataset contains three different scenes, as shown in Section 4.1. By analyzing the data, it is clear that three clusters have formed, as expected. This may happen because different scenes have different complexities, therefore the CAE may have different reconstruction errors for each of them. Figure 8 (bottom) shows the scatter plots of this dataset.

6 CONCLUSION

Automatic video surveillance is an area of growing interest in the computer vision community. Its applications can increase security in public areas such as airports and parks. This work presented a method to detect anomalies in video sequences. Four well known benchmark datasets were used to evaluate the model.

Results have shown that the method can classify anomalous frames with acceptable performance, considering the complexity of the problem. Despite not

Future works will mainly focus on the feature extraction process. Finding the correct attributes is essential to the success of any classifier. Deep Learning methods have shown great performance in the computer vision field, and may be the key to learn meaningful representations from data.

Acknowledgements

M. Gutoski would like to thank CAPES for the scholarship. M. Ribeiro would like to thank the Catarinense Federal Institute of Education, Science and Technology and IFC/CAPES/Prodoutoral for the scholarship. N. Aquino would like to thank the Organization of the American States, the Coimbra Group of Brazilian Universities and the Pan American Health Organization. H.S.Lopes would like to thank to CNPq for the research grant number 440977/2015-0. All authors would like to thank to NVIDIA for the donation of a GPU used in this work.

References

[1] J. Canny. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (6):679–698, 1986.

[2] C. Chen, Y. Shao, and X. Bi. Detection of anomalous crowd behavior based on the acceleration feature. *IEEE Sensors Journal*, 15(12):7252–7261, 2015.

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

[4] B. K. P. Horn and B. G. Schunck. Determining optical flow. *Artificial Intelligence*, 17(1-3):185–203, 1981.

[5] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional Architecture for Fast Feature Embedding. *arXiv:1408.5093*, pages 1–4, 2014.

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

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

[8] J. Masci, U. Meier, D. Cireşan, and J. Schmidhuber. Stacked convolutional auto-encoders for hierarchical feature extraction. In *Proc. 21th Int. Conference on Artificial Neural Networks*, volume I, pages 52–59, 2011.

[9] M. A. F. Pimentel, D. A. Clifton, L. Clifton, and L. Tarassenko. A review of novelty detection. *Signal Processing*, 99:215–249, 2014.

[10] M. Ribeiro, A. E. Lazzaretti, and H. S. Lopes. A study of deep convolutional auto-encoders for anomaly detection in videos. *Pattern Recognition Letters, in press*, 2017.

[11] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning representations by back-propagating errors. *Nature*, 323(1):533–536, 1986.

[12] V. Saligrama and Z. Chen. Video anomaly detection based on local statistical aggregates. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pages 2112–2119, 2012.

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

[14] A. A. Sodemann, M. P. Ross, and B. J. Borghetti. A review of anomaly detection in automated surveillance. *IEEE Transactions on Systems, Man, and Cybernetics, Part C*, 42(6):1257–1272, 2012.

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

[16] D. X., Y. Y., E. R., and N. S. Detecting anomalous events in videos by learning deep representations of appearance and motion. *Computer Vision and Image Understanding*, 156:117–127, 2017. Image and Video Understanding in Big Data.

[17] D. Xu and E. Ricci. Learning deep representations of appearance and motion for anomalous event detection. In *Proc. British Machine Vision Conference*, pages 1–12, 2015.