Two-streams Fully Convolutional Networks for Abnormal Event Detection in Videos

Slim Hamdi[1,2], Samir Bouindour[1], Kais Loukil[2], Hichem Snoussi[1] and Mohamed Abid[2]

[1]*LM2S University of Technology of Troyes, 12, rue Marie Curie - CS 42060, 10004 Troyes cedex, France*

[2]*CES Laboratory, ENIS National Engineering School, University of Sfax, B.P. 3038 Sfax, Tunisia*

Keywords: Deep Learning, Anomaly Detection, Convolutional Auto-encoder.

Abstract: In the context of abnormal event detection in videos, only the normal events are available for the learning process, therefore the implementation of unsupervised learning method becomes paramount. We propose to use a new architecture denoted Two-Stream Fully Convolutional Networks (TS-FCNs) to extract robust representations able to describe the shapes and movements that can occur in a monitored scene. The learned FCNs are obtained by training two Convolutional Auto-Encoders (CAEs) and extracting the encoder part of each of them. The first CAE is trained with sequences of consecutive frames to extract spatio-temporal features. The second is learned to reconstruct optical flow images from the original images, which provides a better description of the movement. We enhance our (TS-FCN) with a Gaussian classifier in order to detect abnormal spatio-temporal events that could present a security risk. Experimental results on challenging dataset USCD Ped2 shows the effectiveness of the proposed method compared to the state-of-the-art in abnormal events detection.

1 INTRODUCTION

Security is a major concern in all modern communities. It helps to strengthen the climate of peace and create a sense of safety conducive to the proper development of the society. The use of video surveillance, which is a recognized tool in the field of security, has become widespread over the last few years. The increase of the video surveillance stream poses the problem of the efficient treatment of the large amounts of data generated. In this context, the automatic detection of abnormal events becomes an important research task and arouses great interest in the scientific community. An abnormal event in video footage is often characterized by abnormal shapes, abnormal movements or their combination. Precursor works have extensively explored the trajectory analysis for the detection of abnormal events (S.Zhou, 2015; C.Piciarelli, 2008). Despite their interest in detecting deviant trajectories in non-crowded scenes, these methods remain sensitive to occlusions in crowded scenes and many of them require efficient tracking algorithms that consume high computing power. Other works proposed to handle these weaknesses using the low-level descriptors such as histograms of oriented gradients (HOG) (V.Mahadevan, 2010) or histogram of optical flow (HOF) (W.Li, 2014). The author in (V.Reddy, 2011) proposed the extraction of multiple characteristics applied on small regions of frames obtained by foreground segmentation technique, a classifier for each feature is then used to pick up the abnormalities. However, methods based on low-level descriptors require prior knowledge to match the features with the corresponded events and they are too local for complex behaviors understanding. Deep Learning (DL) has recently received a significant attention of researchers since it achieved good results in various computer vision tasks. Based on DL technology many works have been proposed to handle the anomaly detection task. (J.Sun, 2017) integrated One-Class Support Vector Machine (OC-SVM) into Convolutional Neural Network (CNN) for end-to-end deep one-class learning framework adapted for anomaly detection. The author of (S.Zhou, 2016) proposed to train a spatio-temporal convolutional neural network on volumes of interest (SVOI) extracted using optical flow to classify events as normal or abnormal. The author of (N.Sebe, 2017) used normal images and corresponding optical-flow representations to train Generative Adversarial Network (GAN). The GAN is then used to detect abnormal shapes and movements based on the principle that normal data should be generated with greater accuracy than abnormal one. (Y.S.Chong, 2017) proposes a network composed of temporal CAE and spatial CAE to extract spatio-temporal features and re-

Hamdi, S., Bouindour, S., Loukil, K., Snoussi, H. and Abid, M.
Two-streams Fully Convolutional Networks for Abnormal Event Detection in Videos.
DOI: 10.5220/0008949405140521
In *Proceedings of the 12th International Conference on Agents and Artificial Intelligence (ICAART 2020) - Volume 2*, pages 514-521
ISBN: 978-989-758-395-7; ISSN: 2184-433X

construct the input data based on these representations. The reconstruction error is then used to dissociate between normal and abnormal frames. In addition to the anomaly detection, our method was inspired by previous works on action recognition, (B.Zhang, 2016) proposed to fuse two convolutional neural networks, one trained on images and the second on Motion Vectors (MV) to extract robust spatio-temporal representations allowing the classification of different actions. Convolutional Auto-Encoders (CAE) is an unsupervised artificial neural network that uses the convolution to learn extracting representations from which the reconstruction of the input image is possible. Based on this concept we propose to introduce in this paper, a new architecture for abnormal event detection . It consists of two fully convolutional networks (FCNs), one formed on images and other on optical flow representations. This combination allows extracting high-level representation able to describe complex behaviors and dissociate between normal and abnormal events.

2 STATE OF THE ART

For many years, the development of a traditional pattern recognition system required extensive expertise and knowledge to extract features from the raw data that could be adapted and used to detect, identify or classify elements among the input data. The abnormal event detection methods that adopted this model inherited the same dependencies. These methods require a priori knowledge to build a feature extractor adapted to the targeted events and the monitored scene. These constraints have favoured the emergence of methods for detecting abnormal events based on the learning of representations and more precisely on deep learning. The learning of representations or the learning of characteristics is a set of methods that automates the step of extracting characteristics. These methods make it possible to define through learning the appropriate transformations to be made to the input data in order to obtain representations that allow a targeted task to be performed such as action recognition, image classification, human pose estimation, semantic segmentation, etc. Deep learning is a subdomain of representation learning, it aims to learn high-level abstractions in data using multi-level architectures. These different levels are obtained by stacking multiple non-linear transformation modules. Each module transforms the data at a different level until an adapted representation is obtained that allows the target task to be performed. Deep learning has overtake the traditional model in some cases of application and it has made the possibility to design effective pattern recognition systems without in-depth expertise on the targeted elements. Convolutional Neural Networks (CNNs) are among the most popular supervised methods of deep learning. This is largely due to the remarkable results obtained with CNNs such as Alexnet, VGG, GoogLeNet and ResNet (K.He, 2016; C.Szegedy, 2015; K.Simonyan, 2014; A.Krizhevsky, 2012) on the international ILSVRC (ImageNet Large-Scale Visual Recognition Challenge). The CNN is a type of artificial neural network whose functioning was inspired by the visual cortex of animals. It consists of several layers that process data hierarchically. The characteristics extracted in the first layers of CNN generally describe the presence of simple shapes (edges and contours), the following layers extract slightly more complex patterns by detecting assemblies of simple shapes while neglecting irrelevant variations. More we explore the network, more the deeper layers describe better complex shapes with an increasing level of abstraction until be able to represent parts of objects or even complete objects in the case of the last layers .

2.1 Related Work

In this way, the CNN amplifies those aspects of the input data that are important for discrimination and removes irrelevant variations. As mentioned above, the CNN is a supervised learning method, to fully exploit these capabilities in terms of feature extraction and classification for anomaly detection, a labelled database containing learning examples from both classes (normal and abnormal) is required. In (C.Ding, 2014), a 3D CNN is proposed to classify video clips into two classes (fight or no fight) in order to detect acts of violence in ice hockey videos. A 3D CNN is characterized by 3D convolutions, which allows it to extract spatio-temporal characteristics essential for the description of motion. In (O.Russakovsky, 2015), a 3D CNN has also been built to classify video volumes of interest SVOI (Spatial-temporal Volumes of Interest) into two classes: normal and abnormal. Volumes of interest are selected through the optical flow, those containing little or no movement are not processed by the CNN. In (R.Hinami, 2017), proposes to combine a multi-tasking Fast R-CNNN with the KDE kernel density estimation method. The multi-task Fast R-CNN multi-task is trained in a supervised way to extract semantic characteristics and classification scores for different objects present in the input images. These characteristics are then used in the KDE to detect anomalies. In this way, not only abnormal events can be detected but also a description of the event detected can be given using the labels provided by the CNN. Dual flow architectures integrating CNNs have also been explored in the context of abnormal event detection. Despite the convincing results of methods based on deep and supervised learning, the need to use both normal and abnormal training examples complicates their integration into intelligent video surveillance systems. It has been demonstrated that a CNN trained to perform a target task can provide generic and robust characteristics, usable to perform another computer vision task for which it has not been specifically trained. (P.Sermanet, 2013; A.Sharif, 2014) prove that a CNN trained only for object classification, can be operated for different tasks such as scene classification, detailed classification, attribute detection, visual instance recovery. The results obtained provide tangible proof of the ability of CNNs to provide generic and robust features that can be used for different computer vision tasks. This principle has been applied in many abnormal event detection projects. In(M.Ravanbakhsh, 2018), a pre-trained CNN is fused with a binary quantization layer whose weights are trained using a binary hashing method called ITQ (Iterative Quantization Hashing)(Y.Gong, 2013). In (M.Sabokrou, 2018b), a pre-trained CNN is combined with a trainable sparse auto-encoder to obtain a two-level characteristic extractors. At the output of the CNN a first Gaussian classifier is used to classify image regions as normal, abnormal or suspicious. The representations of the suspect regions are then transformed by the auto-encoder to obtain more discriminating representations. A second Gaussian classifier is used at the output of the auto-encoder to classify suspect regions into normal and abnormal. Methods based on learning transfer do not require a labelled database for feature extraction and their results in terms of detection and localization are promising. Nevertheless, the dependence of these methods on pre-trained models imposes a certain rigidity on them and considerably reduces their prospects for improvement. These criteria have encouraged the emergence of work oriented towards approaches based on unsupervised learning. The development of learning methods that do not require a labelled database has always been a primary objective in the field of automatic learning. In addition to the difficulty of building labelled databases large enough to capture the complexity of some of the topics covered, this interest in unsupervised learning is inspired in part by the fact that human learning is largely unsupervised (Y.LeCun, 2015). Indeed, man has a considerable capacity to observe, analyse and understand the world around him without using labels for each object. Despite the importance and challenges surrounding this type of learning, the rapid success of the CNN has somewhat eclipsed unsupervised learning for a period of time. Some recent works based on Auto-encoders (AEs) or Sparse coding to extract different linear or non-linear representations of appearance (image) or motion (flow), in order to model normal behaviours in surveillance videos. The AE auto encoder (Auto-Encoder) is a fully connected neural network that is widely used in automatic learning. It consists of an input layer, an output layer and one or more hidden layers. The training of the AE is done by back-propagation of the gradient in order to minimize the reconstruction error between the input and output data. (D.Xu, 2015) proposes AMDN (Appearance and Motion DeepNet) which is a network consisting of three SDAEs (stacked denoising auto-encoders), a first trained to reconstruct patches extracted from normal images, a second trained with the optical flow representations corresponding to the patches and a third trained with the concatenation of the patches and their optical flow representations. Once the three networks have been trained, the representations obtained are used to train three OC-SVMs. A CAE (Convolution Auto Encoder) is an AE with added convolution layers. CAE has been widely explored in the detection of abnormal events.(S.Hamdi, 2019) Abnormal motion is picked by relative thresholding. One-class SVM is trained with spatial features for robust classification of abnormal shapes. Moreover, a decision function is applied to correct the false alarms and the miss detections. (M.Hasan, 2017) proposes two methods also based on CAEs. In the first, the authors suggest a CAE trained to reconstruct low-level characteristics (HOG and HOF) extracted from samples of the normal class. In the second method, they propose to use a spatio-temporal CAE trained directly on video volumes. In both approaches, anomalies are detected thanks to a regularity score calculated with the reconstruction error. (Y.H.Tay, 2017) proposes to use the reconstruction error of a spatio-temporal CAE to detect abnormal events. The proposed CAE integrates 2D convolution layers for learning spatial characteristics and ConvLSTMs (convolutional long short term memory) for temporal characteristics. In recent years, the use of GANs (Generative Adversarial Networks) has increased considerably in the field of automatic learning. GAN is an unsupervised learning algorithm initially proposed by (I.Goodfellow, 2014). It consists of two sub-networks, a generator and a discriminator placed in competition. During the learning phase the generator tries to generate convincing data to lure the discriminator, who tries to detect whether the data is real or generated. In this way we obtain two trained networks, one to generate realistic data and the other to distinguish real data. After the learning phase, the generator can be used independently to create data (Y.Jin, 2017; P.Isola, 2017), or for discrimination tasks (W.Liu, 2018; M.Ravanbakhsh, 2017). but it can also be used in conjunction with the discriminator (S.Xingjian, 2015). (M.Sabokrou, 2018a) offers a method called AVID (Adversarial Visual Irregularity Detection) to detect and locate irregularities in videos. A GAN composed of a generator trained to remove irregularities in the input images and replace them with the dominant patterns of the same images and a discriminator in the form of an FCN that predicts the probability that the different regions (patches) of the input images will be abnormal. The two networks are trained in an adversarial manner and the irregularities are simulated using Gaussian noise. After the learning phase, each of the two networks is able to detect irregularities: the generator at the pixel level thanks to the error between the original and generated images, the generator has been trained to erase the irregularities, so when an image containing irregularities is introduced, the generator eliminates these irregularities and replaces them with other reasons which will result in a larger generation error. The discriminator, on the other hand, can directly predict the probability of a patch containing irregularities.

3 PROPOSED METHODS

Recently deeper two-streams convolutional networks have been applied successfully on action recognition. Based on this concept we propose a new efficient architecture composed of two FCNs to tackle the problem of anomaly detection in video into both different methods.

3.1 TS-FCN 1

Our proposed architecture consists of two parts: spatio-temporal FCN (ST-FCN) for learning representations from video frames, and optical flow FCN (OF-FCN) to strengthen the movement description of the learned representations. The learned two FCNs are obtained by training two convolutional auto-encoders (CAEs) in order to reconstruct video volumes and extracting the encoder part of each of them, Figure (1).

The spatio-temporal CAE and the optical flow CAE are respectively learned using normal training samples and corresponding optical-flow representations. Both CAEs have the same architecture and each of them is composed by four 3D convolution layers (encoder) and four 3D deconvolution layers (decoder) . The convolution layers encode representations from the input data while the deconvolution reflect the encoder part to reconstruct them. The spatio-temporal CAE takes as input 3D volumes of three consecutive frames F: $\{F_t;\ F_{t-1};\ F_{t-2}\}$. The optical flow CAE, 3D volumes of three optical flow representations OF: $\{OF_t;\ OF_{t-1};\ OF_{t-2}\}$, where OF_t is obtained by extracting the optical flow for each two consecutive frames Figure (2). After training the CAEs, the encoder part of each of them represent the FCNs of our two-stream architecture. For each input frame F_t represented by the video volumes F and OF, each network provides a feature map of dimension 676*256. We combine these two features maps to obtain representation of dimension 676*512, where each row (feature vector) represents a patch of size 27*27 of the original input frame. This architecture allows us, by means of the first FCN, to obtain a robust spatio-temporal representation of each patch of the input frame and refine this representation using the second FCN, which allows a more robust representation of the movement by using the optical flow descriptor. Thanks to this architecture, each small region of the input video volumes is represented by a feature vector able to describe the shapes and the movements contained in this region. In test phase both optical flow and original frames are used and we propose to complete our architecture with a robust

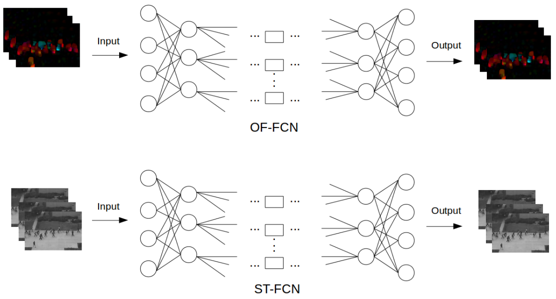

Figure 1: Our TS-FCN 1 Architecture.

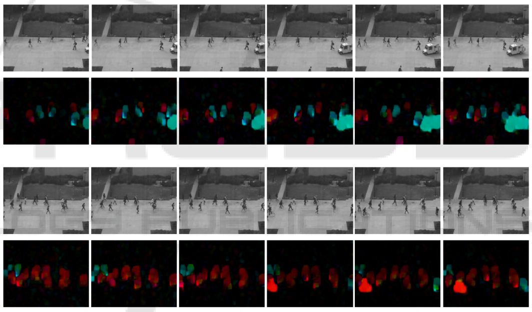

Figure 2: Optical flow and original images.

Gaussian classifier that allows us to dissociate between the normal and abnormal patches of each frame through the classification of their representative feature vectors. The classification of the feature vectors corresponding to small regions of the input images is carried out according to the following process: in the first step, we extract feature vectors $X = \{x_i\}, x_i \in R^{512}$ from the normal training examples, the mean M and the inverse of the covariance matrix Q of X are then calculated. In the second step, we evaluate each feature vector x_j of the testing frames with Mahalanobis distance d_j using M and Q. This is represented in the following equation:

$$d_j = (x_j - M) * Q * (x_j - M)' \qquad (1)$$

The outlier vectors, which actually represents abnormal frames, are then picked by thresholding the distance. If the distance exceeds a threshold α, the vector x_j is considered as outlier and the frame p_j is labeled as abnormal, Eq (2).

$$p_j : \begin{cases} Normal & if \quad d_j \leq \alpha \\ Abnormal & if \quad d_j > \alpha \end{cases} \qquad (2)$$

Table 1: CAEs parameters.

Layer	Filters	Kernel (h,w,d)	Stride(h,w,d)
Conv1	64	[11,11,1]	[2,2,1]
Conv2	96	[3,3,1]	[1,1,1]
Conv3	128	[3,3,3]	[2,2,1]
Conv4	256	[3,3,1]	[2,2,1]
Deconv1	256	[3,3,1]	[2,2,1]
Deconv2	128	[3,3,3]	[2,2,1]
Deconv3	96	[3,3,1]	[1,1,1]
Deconv4	1	[11,11,1]	[2,2,1]

3.2 TS-FCN 2

The extraction of optical flow images in the test phase allows the system to execute an additional task to extract optical flow images. Moreover, in the training phase the representations of the two volumes of two streams are independents. Then we propose a second method based on new architecture of one block to represent our TS-FCN to rectify the imperfections of the first method. This TS-FCN is learned using normal training samples representations of original images only. It is composed by eight 3D convolution layers (encoder) , eight 3D deconvolution layers (decoder) and one concatenation layer to combine both presentations. The TS-FCN takes as input 3D volumes of three consecutive frames F: $\{F_t; F_{t-1}; F_{t-2}\}$. and try not only to reconstruct those frames but also to reconstruct the optical flow 3D volumes of OF: $\{OF_t; OF_{t-1}; OF_{t-2}\}$ at the same time. The Mean Squared Error is used as loss function to train our model (figure 3). After training phase, the encoder part contained 8 convolutions layers represents our TS-FCN. Our model provides a feature map of dimension 676*512 in this case. It is capable to obtain a robust spatio-temporal representation of each both shapes and motion into frames. In test phase, we do not need to extract the optical flow represention manually but our model is capable to construct new representation of optical flow from original frames which are more dedicated to the task of the detection of anomalies (figure 4). However, the classification task is done as the same way of the first method.

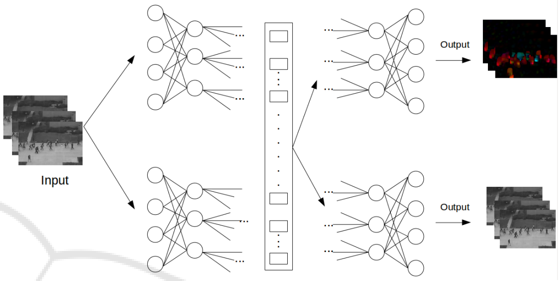

Figure 3: Our TS-FCN 2 Architecture.

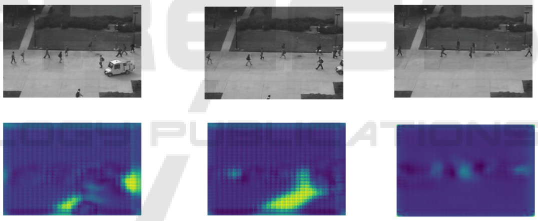

Figure 4: Constructed optical flow in TS-FCN 2.

4 EXPERIMENT RESULTS

To evaluate the proposed architecture, we used the USCD Ped2 dataset and compared our results to the state-of-the-art methods. The UCSD Ped2 dataset has 16 folders of training and 12 for testing. The training part of the dataset contains only normal events summarized in pedestrian movements. The testing folders in addition to pedestrians also contain abnormal events that result in the appearance of non-pedestrians. We evaluate our different methods using (Error Equal Rate) EER and (Area Under Curve ROC) AUC as evaluations criteria. A smaller EER corresponds with better performance. As for the AUC, a bigger value corresponds with better performance. The frames and their corresponding optical flow representations are extracted from the raw videos and resized to 227×227. We then subtract a mean image from each frame contained in the same folder. The mean image is obtained by averaging the frames of each training folder. After the mean subtraction, we scale the pixel values between -1 and 1. For the testing images, we use the mean image calculated during the training to ensure the condition of real world applications. We then group these pre-processed images and the optical flow representations in video volumes composed of 3 consecutive frames. During the training procedure these video volumes are then introduced as inputs to train the two CAE (method 1). We train the two CAE by minimizing the reconstruction error of the input volumes using Adam optimizer. A hyperbolic tangent is used as activation function of each convolution and deconvolution layer to ensure the symmetry of the reconstructed and the input video volumes. The detailed parameters of our network are provided in Table [1]. However only group of pre-processed images is used to train our architecture to reconstruct both pre-processed images and the optical flow representations (method 2). During the testing phase we use only the encoder parts (FCNs), with Gaussian classifier to detect abnormalities in the testing frames. A comparison with state-of-the-art methods are related in Table [2]. We evaluate not only our both methods but also the spatio-temporal FCN (ST-FCN) individually. These experiments demonstrate the utility of combining the two FCNs as the EER progresses from 19% to 13% (method 1). Which make the importance of using of optical flow image to represent the motion in each frames. Moreover method 2 proves that the coherence of both shapes and motion features in the training phase makes our architecture more robust. It obtained an EER egal to 8.45% and achieves AUC more than 93%. The ROC curve is plotted according to the detection results. The FPR is the rate of incorrectly detected frames to all normal frames in ground truth and the TPR is the rate of correctly detected frames to all abnormal frames in ground truth. We quantify the performance in terms of the equal error rate (EER) and the area under ROC curve (AUC). The EER is the point on the ROC curve that FPR is equal to (1-TPR). Our two-stream fully convolutional networks combined with simple classifier demonstrates good performances, equivalent with state-of-the-art methods for anomaly detection detection.

Table 2: EER and AUC for frame level comparisons on ped2 dataset

Methods	EER	AUC
PCA(D.-S. Pham, 2011)	29.20	73.98
CAE(FR)(M. Ribeiro, 2017)	26.00	81.4
ConvAE(M. Hasan, 2016)	21.7	90.00
EAD(Hung Vu, 2018)	16.47	86.43
Chong(Chong and Tay, 2017)	12	-
Sabokrou(M. Sabokrou, 2017)	8.2	-
ours		
ST-FCN	19	87.15
TS-FCN 1	**13.2**	**91.6**
TS-FCN 2	**8.45**	**93.6**

5 CONCLUSION

In this paper, a new unsupervised methods were proposed to train FCNs. We used these methods to learn a new architectures composed of two FCNs, one trained on video volumes and the second on optical flow representations. Our two-stream fully convolutional networks allows extracting high level spatio-temporal features taking into account the movements and shapes present in each small region of the video. This robust representation makes possible, with a simple classifier, to differentiate between normal and abnormal events. We have tested our TS-FCN on challenging dataset, containing crowded scenes (USCD Ped2) Our method obtained high results competing the best state-of-the-art methods in detection of abnormal events.

Our future works will investigate the strengthening of our learning process by adding a custom loss function. This will ensure not only the good quality of the features by aiming at the reconstruction of input data but will also guarantee the compactness of the representations of the data belonging to the same class. This will allow to efficiently dissociate between normal and abnormal events.

REFERENCES

A.Krizhevsky, I. G. (2012). Imagenet classification with deep convolutional neural networks. in Advances in neural information processing systems ,pp.1097#1105,.

A.Sharif, R. H. J. S. (2014). Cnn features off-the-shelf:an astounding base line for recognition. in Proceedings of the IEEE conference on computer vision and pattern recognition workshops,pp.806#813.

B.Zhang, L. Z. Y. H. (2016). Real-timeaction recognition with enhanced motion vector cnns. in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition pp. 2718–2726,.

C.Ding, S. M. W. B. (2014). Violence detection in video by using 3d convolutional neural networks. in International Symposium on Visual Computing ,pp.551#558, Springer.

Chong, Y. S. and Tay, Y. H. (2017). Abnormal event detection in videos using spatiotemporal autoencoder. inInternational Symposium on Neural Networks, pp. 189–196, Springer.

C.Piciarelli, C. G. (2008). a trajectory-based anomalous event detection,. Trajectory-based anomalous event detection,"IEEE Transactions on Circuits and Systems for video Technology, vol. 18, no. 11, pp. 1544–1554,.

C.Szegedy, W. Y. P. S. D. D. V. A. (2015). Going deeper with convolutions. in Proceedings of the IEEE confe-rence on computer vision and pattern recognition,pp.1#9.

D.-S. Pham, B. S. D. Q. P. S. V. (2011). Detection of cross-channel anomalies from multiple data channels. nICDM,2011, pp. 527–536.

D.Xu, E. Y. J. N. (2015). Learning deep representations of appearance and motion for anomalous event detection. arXiv preprint arXiv:1510.01553.

Hung Vu, T. D. N. D. P. (2018). Detection of unknown anomalies in streaming videos with generative energy-based boltzmann models. arXiv preprint arXiv:1805.01090.

I.Goodfellow, J.-A. M. B. D.-F. S. A. Y. (2014). Generative adversarial nets. in Advances in neural information processing systems ,pp.2672#2680.

J.Sun, J. C. H. (2017). Abnormal event detection for video surveillance using deep one-class learning. Multimedia Tools and Applications, pp. 1–15,.

K.He, X. S. J. (2016). Deep residual learning for image recognition. in Proceedings of the IEEE conference on computer vision and pattern recognition,pp.770#778.

K.Simonyan, A. (2014). Very deep convolutional networks for large-scale image recognition. ar Xiv preprint arXiv:1409.1556.

M. Hasan, J. C. J. N. A. K. R.-C. L. S. D. (2016). "learning temporal regularity in video sequences,. in CVPR, 2016.

M. Ribeiro, A. E. L. H. S. L. (2017). A study of deep convolutional auto-encoders for anomaly detection in videos. Pattern Recognition Letters.

M. Sabokrou, M. F. M. F. R. K. (2017). Deep-cascade:cascading 3d deep neural networks for fast anomaly detection and localization in crowded scenes. IEEE Transactions on Image Processing, vol. 26, no. 4, pp. 1992–2004.

M.Hasan, J. J. A.-C. L. (2017). Learning temporal regularity in video sequences. in Proceedings of the IEEE conference on computer vision and pattern recognition,pp.733#742.

M.Ravanbakhsh, M. E. L. C. N. (2017). Abnormal event detection in videos using generative adversarial nets. in 2017 IEEE International Conferenceon Image Processing (ICIP),pp.1577#1581,IEEE, 2017.

M.Ravanbakhsh, M. H. E. N. S. Z. A. A. M. a. (2018). Plug-and-play cnn for crowd motion analysis:an application in abnormal event detection. in 2018 IEEE Winter Conference on Applications of Computer Vision (WACV),pp.1689#1698,IEEE.

M.Sabokrou, M. M. R. M. J. E. (2018a). Avid: Adversarial visual irregularity detection. arXiv preprint arXiv:1805.09521.

M.Sabokrou, M. M. Z. R. (2018b). Deep-anomaly :fully convolutional neural network for fast anomaly detection in crowded scenes. Computer Vision and Image Understanding,vol.172,pp.88#97.

N.Sebe, M. M. E. L. M. C. R. (2017). Abnormal event detection in videos using generative adversarial nets. in Image Processing (ICIP), 2017 IEEE International-Conference on, pp. 1577–1581, IEEE,.

O.Russakovsky, J. H. J. S. S. Z. A. A. M. a. (2015). Imagenet large scale visual recognition challenge. International journal of computer vision ,vol.115,no.3,pp.211#252.

P.Isola, J.-Y. T. A. (2017). Image-to-image translation with conditional adversarial networks. n Proceedings of the IEEE conference on computer vision and pattern recognition ,pp.1125#1134.

P.Sermanet, D. X. M. R. Y. (2013). Overfeat :integrated recognition ,localization and detection using convolutional networks. arXiv preprint arXiv:1312.6229.

R.Hinami, T. S. (2017). Joint detection and recounting of abnormal events by learning deep generic knowledge. in Proceedings of the IEEE International Conference on Computer Vision,pp.3619#3627.

S.Hamdi, S. K. H. M. (2019). Hybrid deep learning and hof for anomaly detection. 2019 6th International Conference on Control, Decision and Information Technologies (CoDIT), IEEE.

S.Xingjian, Z. H. D. W. W. (2015). Convolutional lstm network :a machine learning approach for precipitation now casting. in Advances in neural information processing systems,pp.802#810.

S.Zhou, W. D. Z. (2015). Unusual event detection in crowded scenes by trajectory analysis. in Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on,pp. 1300–1304, IEEE,.

S.Zhou, W.Shen D.Zeng M.Fang, Y. Z. (2016). Spatial–temporal convolutional neural networks for anomaly detection and localization in crowded scenes. Signal Processing: Image Communication, vol. 47, pp. 358–368.

V.Mahadevan, W. V. N. (2010). Anomaly detection in crowded scenes. in Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, pp. 1975–1981, IEEE.

V.Reddy, C. B. (2011). Improved anomaly detectionin crowded scenes via cell-based analysis of foreground speed, sizeand texture. in Computer Vision and Pattern Recognition Workshops(CVPRW), 2011 IEEE Computer Society Conference on, pp. 55–61,IEEE.

W.Li, V. N. (2014). Anomaly detection and localization in crowded scenes,. IEEE transactions on pattern analysis and machine intelligence, vol. 36, no. 1, pp. 18–32.

W.Liu, W. D. S. (2018). Future frame prediction for anomaly detection a new base line. in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition ,pp.6536#6545.

Y.Gong, S. A. F. (2013). A procrustean approach to learning binary codes for large-scale image retrieval. IEEE transactions on pattern analysis and machine intelligence ,vol.35,no.12,pp.2916#2929.

Y.H.Tay, Y. (2017). Abnormal event detection in videos using spatio-temporal autoencoder. in International Symposiumon Neural Networks ,pp.189#196,Springer.

Y.Jin, J. M. Y. H. Z. (2017). Towards the automatic anime characters creation with generative adversarial networks. arXiv preprint arXiv:1708.05509.

Y.LeCun, Y. G. (2015). Deep learning. nature, vol.521, no.7553,p.436.

Y.S.Chong, Y. (2017). Abnormal event detection in videos using spatiotemporal auto-encoder. in International Symposium on Neural Networks, pp. 189–196, Springer.