ABNORMAL EVENT DETECTION IN VIDEOS USING GENERATIVE ADVERSARIAL NETS

Mahdyar Ravanbakhsh [1], *Moin Nabi* [2], *Enver Sangineto* [2], *Lucio Marcenaro*[1], *Carlo Regazzoni* *[1,3], *Nicu Sebe* [2]

[1] DITEN, University of Genova [2] DISI, University of Trento [3] Carlos III University of Madrid

ABSTRACT

In this paper we address the abnormality detection problem in crowded scenes. We propose to use Generative Adversarial Nets (GANs), which are trained using *normal* frames and corresponding optical-flow images in order to learn an internal representation of the scene *normality*. Since our GANs are trained with only normal data, they are not able to generate abnormal events. At testing time the real data are compared with both the appearance and the motion representations reconstructed by our GANs and abnormal areas are detected by computing local differences. Experimental results on challenging abnormality detection datasets show the superiority of the proposed method compared to the state of the art in both frame-level and pixel-level abnormality detection tasks.

Index Terms— Video analysis, abnormal event detection, crowd behaviour analysis, Generative Adversarial Networks

1. INTRODUCTION

Abnormality detection in crowds is motivated by the increasing interest in video-surveillance systems for public safety. However, despite a lot of research has been done in this area in the past years [1, 2, 3, 4, 5, 6, 7], the problem is still open.

There are two main reasons for which abnormality detection is challenging. First, existing datasets with *ground truth* abnormality samples are small. This limitation is particularly significant for deep-learning based methods, which have shown an impressive accuracy boost in many other recognition tasks [8, 9, 10, 11, 12, 13] but are data-hungry. The second reason is the lack of a clear and objective definition of abnormality. Moreover, these two problems are related to each other, because the abnormality definition subjectivity makes it harder to collect abnormality ground truth.

In order to deal with these problems, *generative* methods for abnormality detection focus on modeling only the *normal* pattern of the crowd. The advantage of the generative paradigm lies in the fact that only *normal* samples are needed at training time, while detection of what is abnormal is based on measuring the distance from the learned normal pattern. However, most of the existing generative approaches rely on hand-crafted features to represent visual information

*Carlo Regazzoni has contributed to produce this work partially under the program "UC3M-Santander Chairs of Excellence".

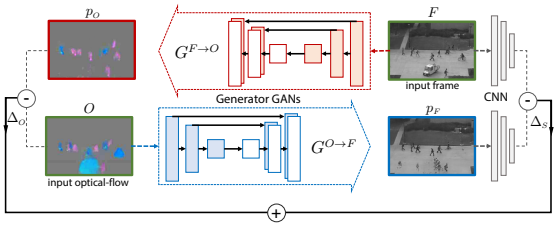


Fig. 1. Top: a generator network takes as input a frame and produces a corresponding optical-flow image. Bottom: a second generator network is fed with a real optical-flow image and outputs an appearance reconstruction.

[4, 14, 3, 7, 2] or use Convolutional Neural Networks (CNNs) trained on external datasets [15, 16]. Recently, Xu et al. [17] proposed to use stacked denoising autoencoders. However, the networks used in their work are shallow and based on small image patches. Moreover, additional one-class SVMs need to be trained on top of the learned representation.

In this paper we propose a generative deep learning method applied to abnormality detection in crowd analysis. More specifically, our goal is to use deep networks to learn a representation of the *normal pattern* utilizing only *normal* training samples, which are much easier to collect. For this purpose, Generative Adversarial Networks (GANs) [18] are used, an emerging approach for training deep networks using only unsupervised data. While GANs are usually used to generate images, we propose to use GANs *to learn the normality of the crowd behaviour*. At testing time the trained networks are used to generate appearance and motion information. Since our networks have learned to generate *only* what is normal, they are not able to reconstruct appearance and motion information of the possible *abnormal* regions of the test frame. Exploiting this intuition, a simple difference between the real test-frame representations and the generated descriptions allows us to easily and robustly detect abnormal areas in the frame. Extensive experiments on challenging abnormality detection datasets show the superiority of the proposed approach compared to the state of the art.

2. BACKGROUND

Abnormality Detection Our method is different from [4, 14, 3, 7, 2, 19, 20, 6, 5, 21, 22, 23, 24], which also focus on

learning generative models on motion and/or appearance features. A key difference compared to these methods is that they employ hand-crafted features (e.g., Optical-flow, Tracklets, etc.) to model normal-activity patterns, whereas our method learns features from raw-pixels using a deep learning based approach. A deep learning-based approach has been investigated also in [15, 16]. Nevertheless, these works use existing CNN models trained for other tasks (e.g., object recognition) which are adapted to the abnormality detection task. For instance, Ravanbakhsh et al. [15] propose a Binary Quantization Layer plugged as a final layer on top of a CNN, capturing temporal motion patterns in video frames for the task of abnormality segmentation. Differently from [15], we specifically propose to train a deep generative network *directly* for the task of abnormality detection.

Most related to our paper is the work of Xu et al. [17], who propose to learn motion/appearance feature representations using stacked denoising autoencoders. The networks used in their work are relatively shallow, since training deep autoencoders on small abnormality datasets is prone to overfitting. Moreover, their networks are not end-to-end trained and the learned representation need externally trained classifiers (multiple one-class SVMs) which are not optimized for the learned features. Conversely, we propose to use adversarial training for our representation learning. Intuitively, the adopted conditional GANs provide data augmentation and implicit data supervision thank to the discriminator network. As a result we can train much deeper generative networks on the same small abnormality datasets and we do not need to train external classifiers.

GANs [18, 25, 26] are based on a two-player game between two different networks, both trained with unsupervised data. One network is the *generator* (G), which aims at generating realistic data (e.g., images). The second network is the *discriminator* (D), which aims at discriminating real data from data generated from G. Specifically, the *conditional* GANs [18], that we use in our approach, take as input an image x and generate a new image p. D tries to distinguish x from p, while G tries to "fool" D producing more and more realistic images which are hard to be distinguished. Very recently Isola et al. [27] proposed an "image-to-image translation" framework based on conditional GANs, where both G and D are conditioned on the real data. They show that a U-net encoder-decoder with skip connections can be used as the generator architecture together with a patch-based discriminator in order to transform images with respect to different representations. A similar framework is adopted here, generating optical-flow images from raw-pixel frames and vice versa. However, we do not aim at generating images which look realistic, but we use G to learn the normal pattern of an observed crowd scene. At testing time, G is used to generate appearance and motion information of the normal content of the input frame. Comparing this generated content with the real frame allows us to detect the possible abnormal areas of the frame.

3. LEARNING THE NORMAL CROWD BEHAVIOUR

We use the framework proposed by Isola et al. [27] to learn the normal behaviour of the observed scene. Specifically, let F_t be the t-th frame of a training video and O_t the optical-flow obtained using F_t and F_{t+1}. O_t is computed using [28]. We train two networks: $\mathcal{N}^{F\rightarrow O}$, which generates optical-flow from frames and $\mathcal{N}^{O\rightarrow F}$, which generates frames from optical-flow. In both cases, inspired by [27], our networks are composed of a conditional generator G and a conditional discriminator D (we refer to [27] for the architectural details of G and D). G takes as input an image x and a noise vector z (drawn from a noise distribution $\mathcal{Z}$) and outputs an image $p = G(x, z)$ of the same dimensions of x but represented in a different channel. For instance, in case of $\mathcal{N}^{F\rightarrow O}$, x is a frame ($x = F_t$) and p is *the reconstruction* of its corresponding optical-flow image $y = O_t$. On the other hand, D takes as input two images (either (x, y) or (x, p)) and outputs a scalar representing the probability that both its input images came from the real data.

G and D are trained using both a conditional GAN loss $\mathcal{L}_{cGAN}$ and a reconstruction loss $\mathcal{L}_{L1}$. In case of $\mathcal{N}^{F\rightarrow O}$, the training set is composed of pairs of frame-optical flow images $\mathcal{X} = \{(F_t, O_t)\}$, where O_t is represented using a standard three-channels representation of the horizontal, the vertical and the magnitude components. $\mathcal{L}_{L1}$ is given by:

$$\mathcal{L}_{L1}(x, y) = ||y - G(x, z)||_1 \tag{1}$$

while the conditional adversarial loss $\mathcal{L}_{cGAN}$ is:

$$\mathcal{L}_{cGAN}(G, D) = \mathbb{E}_{(x,y)\in\mathcal{X}}[\log D(x, y)]+ \tag{2}$$

$$\mathbb{E}_{x\in\{F_t\},z\in\mathcal{Z}}[\log(1 - D(x, G(x, z)))] \tag{3}$$

Conversely, in case of $\mathcal{N}^{O\rightarrow F}$, we use $\mathcal{X} = \{(O_t, F_t)\}$. We refer to [27] for more details about the training procedure. What is important to highlight here is that both $\{F_t\}$ and $\{O_t\}$ are collected using the frames of the only *normal* videos of the training dataset. The fact that we do not need videos showing abnormal events at training time makes it possible to train our networks with potentially very large datasets without the need of ground truth samples describing abnormality.

At testing time we use only the generators ($G^{F\rightarrow O}$ and $G^{O\rightarrow F}$) corresponding to the trained networks. Since $G^{F\rightarrow O}$ and $G^{O\rightarrow F}$ have observed only normal scenes during training, they are not able to reconstruct an abnormal event. For instance, in Fig. 1 (top) a frame F, containing a vehicle unusually moving on a University campus, is input to $G^{F\rightarrow O}$ and in the generated optical flow image (p_O) the abnormal area corresponding to that vehicle is not correctly reconstructed. Similarly, when the real optical flow (O) associated with F is input to $G^{O\rightarrow F}$, the network tries to reconstruct the area corresponding to the vehicle but the output is a set of unstructured blobs (Fig. 1, bottom). We exploit this *inability* of our networks to reliably reconstruct abnormality to detect possible anomalies as explained in the next section.

4. ABNORMALITY DETECTION

At testing time we input $G^{F\to O}$ and $G^{O\to F}$ using each frame F of the test video and its corresponding optical-flow image O, respectively. Note that the random noise vector z is internally produced by the two networks using dropout [27], and in the following we drop z to simplify our notation. Using F, an optical-flow reconstruction can be obtained: $p_O = G^{F\to O}(F)$, which is compared with O using a simple pixel-by-pixel difference, obtaining $\Delta_O = O - p_O$ (see Fig. 1). Δ_O highlights the (local) differences between the real optical flow and its reconstruction and these differences are higher in correspondence of those areas in which $G^{F\to O}$ was not able to generate the abnormal behaviour.

Similarly, we obtain the appearance reconstruction $p_F = G^{O\to F}(O)$. As shown in Fig. 1 (bottom), the network generates "blobs" in the abnormal areas of p_F. Even if these blobs have an appearance completely different from the corresponding area in the real image F, we empirically observed that a simple pixel-by-pixel difference between F and p_F is less informative than the difference computed in the optical-flow channel. For this reason, a "semantic" difference is computed using another network, pre-trained on ImageNet [29]. Specifically, we use AlexNet [8]. Note that AlexNet is trained using supervised data which are pairs of images and object-labels contained in ImageNet. However, no supervision about crowd abnormal behaviour is contained in ImageNet and the network is trained to recognize generic objects. Let $h(F)$ be the $conv_5$ representation of F in this network and $h(p_F)$ the corresponding representation of the appearance reconstruction. The fifth convolutional layer of AlexNet (before pooling) is chosen because it represents the input information in a sufficiently abstract space and is the last layer preserving geometric information. We can now compute a semantics-based difference between F and p_F: $\Delta_S = h(F) - h(p_F)$.

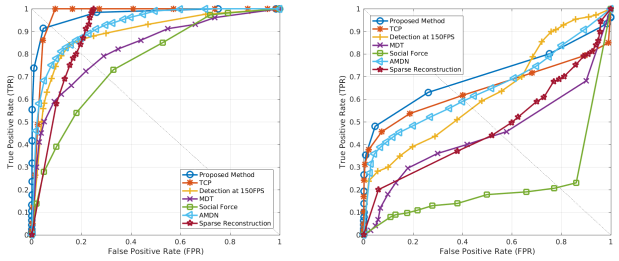
Finally, Δ_S and Δ_O are fused in order to obtain a unique abnormality map. Specifically, we first upsample Δ_S in order to obtain Δ'_S with the same resolution as Δ_O. Then, both Δ'_S and Δ_O are normalized with respect to their corresponding channel-value range as follows. For each test video V we compute the maximum value m_O of all the elements of Δ_O over all the input frames of V. The normalized optical-flow difference map is given by:

$$N_O(i,j) = 1/m_O \Delta_O(i,j). \tag{4}$$

Similarly, the normalized semantic difference map N_S is obtained using m_S computed over all the elements of Δ'_S in all the frames of V:

$$N_S(i,j) = 1/m_S \Delta'_S(i,j). \tag{5}$$

The final abnormality map is obtained by summing N_S and N_O: $A = N_S + \lambda N_O$. In all our experiments we use $\lambda = 2$. A is our final abnormality heatmap.



(a) Frame-level ROC curves. (b) Pixel-level ROC curves.

Fig. 2. ROC curves on Ped1 (UCSD dataset).

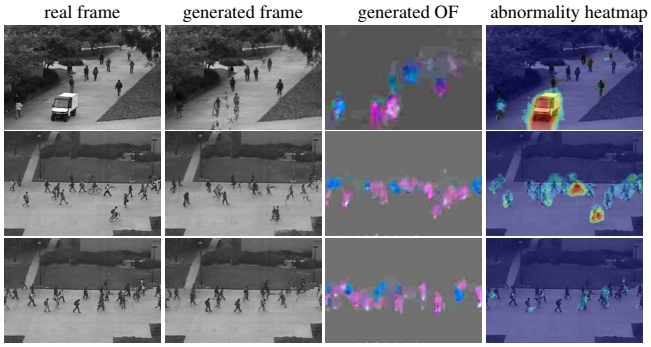


Fig. 3. Some examples of abnormality localization on UCSD.

5. EXPERIMENTAL RESULTS

In this section we evaluate our method using two well-known crowd abnormality datasets. We use both a *pixel-level* and a *frame-level* protocol under the original evaluation setup [1]. The rest of this section describes the datasets, the experimental setup and the obtained results.

GANs Setup. In our experiments, $\mathcal{N}^{F\to O}$ and $\mathcal{N}^{O\to F}$ are trained with the train sequences of the UCSD dataset. All frames are resized to 256×256 pixels. Training is based on stochastic gradient descent with momentum 0.5, batch size 1. Each network is trained for 10 epochs.

Datasets and Experimental Setup. We use two standard datasets: the UCSD Anomaly Detection Dataset [3] and the UMN SocialForce [4]. The **UCSD dataset** is split into two subsets: *Ped1*, which contains 34 train and 16 test sequences, and *Ped2*, which contains 16 train and 12 test videos. This dataset is challenging due to the low-resolution images, different types of moving objects, the presence of one or more anomalies in the scene. The **UMN dataset** contains 11 videos in 3 different scenes, with a total amount of 7700 frames.

5.1. Results and Discussion

Frame-level abnormality detection. The frame-level abnormality detection criterion is based on checking if the frame

Method	Ped1 (frame-level)		Ped1 (pixel-level)		Ped2 (frame-level)	
	EER	AUC	EER	AUC	EER	AUC
MPPCA [2]	40%	59.0%	81%	20.5%	30%	69.3%
Social force(SF) [4]	31%	67.5%	79%	19.7%	42%	55.6%
SF+MPPCA [3]	32%	68.8%	71%	21.3%	36%	61.3%
SR [7]	19%	—	54%	45.3%	—	—
MDT [3]	25%	81.8%	58%	44.1%	25%	82.9%
Detection at 150fps [5]	15%	91.8%	43%	63.8%	—	—
Plug-and-Play CNN [15]	8%	95.7%	40.8%	64.5%	18%	88.4%
AMDN (double fusion) [17]	16%	92.1%	40.1%	67.2%	17%	90.8%
Proposed Method	**8%**	**97.4%**	**35%**	**70.3%**	**14%**	**93.5%**

Table 1. Comparison with the state of the art on the UCSD dataset. The values of the other methods are taken from [17].

Method	AUC
optical-flow [4]	0.84
SFM [4]	0.96
Sparse Reconstruction [7]	0.97
Commotion [30]	0.98
Plug-and-Play CNN [15]	0.98
Proposed Method	**0.99**

Table 2. Results on the UMN dataset (all but our values are taken from [30]).

contains at least one predicted abnormal pixel: in this case the abnormal label is assigned to the whole frame. The procedure is applied over a range of thresholds to build a ROC curve. We compare our method with the state of the art. Quantitative results using both EER (Equal Error Rate) and AUC (Area Under Curve) are shown in Tab. 1, and the ROC curves in Fig. 2. The proposed method is also evaluated on UMN dataset using the same frame level evaluation (Tab. 2).

Pixel-level abnormality localization. The goal of the pixel-level evaluation is to measure the accuracy of the abnormality *localization*. Following [1], a true positive prediction should cover at least 40% the ground truth abnormal pixels, otherwise the frame is counted as a false positive. Fig. 2 shows the ROC curves of the localization accuracy over USDC, and Tab. 1 reports a quantitative comparison with the state of the art. The results reported in Tab. 1-2 show that the proposed approach sharply overcomes all the other compared methods.

Information fusion analysis. In order to analyze the impact on the accuracy provided by each network, $\mathcal{N}^{O\to F}$ and $\mathcal{N}^{F\to O}$, we perform a set of experiments on UCSD Ped1. In the frame-level evaluation, $\mathcal{N}^{O\to F}$ obtains 84.1% AUC and $\mathcal{N}^{F\to O}$ 95.3% AUC, which are lower than the 97.4% obtained by the fused version. In the pixel-level evaluation, however, the performance of $\mathcal{N}^{O\to F}$ dropped to 30.1%, while the $\mathcal{N}^{F\to O}$ is 66.2%. We believe this is due to the low resolution of Δ_S (computed over the results obtained using $\mathcal{N}^{O\to F}$), which makes the pixel-level localization a hard task. By fusing appearance and motion we can refine the detected area, which leads to a better localization accuracy.

Qualitative results. Fig. 3 shows some results using the standard visualization protocol for abnormality localization (red pixels represent abnormal areas). The figure shows that our approach can successfully localize different abnormality types. Moreover, since the generator learned a spatial distribution of the normal motion in the scene, common perspective issues are automatically alleviated. Fig. 3 also shows the intuition behind our approach. Normal objects and events (e.g., walking pedestrians) are generated with a sufficient accuracy. However, the generators are not able to reproduce abnormal objects and events (e.g., a vehicle in the first row) and this inability in reproducing abnormalities is what we exploit in order to detect abnormal areas.

The last row in Fig. 3 shows a failure case, miss detecting the abnormal object (a skateboard). The failure is probably due to the fact that the skateboard is very small, has a "normal" motion (the same speed of normal pedestrians), and is partially occluded.

6. CONCLUSIONS

In this paper we addressed the problem of abnormality detection in crowd videos. We proposed a generative deep learning method based on two conditional GANs. Since our GANs are trained using only normal data, they are not able to generate abnormal events. At testing time, a local difference between the real and the generated images is used to detect possible abnormalities. Experimental results on standard datasets show that our approach outperforms the state of the art with respect to both the frame-level and the pixel-level evaluation protocols. As future work we will investigate the use of Dynamic Images [31] as an alternative to optical-flow in order to represent motion information collected from more than one frame, as suggested by an anonymous reviewer of this paper.

7. REFERENCES

[1] W. Li, V. Mahadevan, and N. Vasconcelos, "Anomaly detection and localization in crowded scenes," *PAMI*, 2014.

[2] J. Kim and K. Grauman, "Observe locally, infer globally: a space-time mrf for detecting abnormal activities with incremental updates," in *CVPR*, 2009.

[3] V. Mahadevan, W. Li, and N. Vasconcelos, "Anomaly detection in crowded scenes," in *CVPR*, 2010.

[4] R. Mehran, A. Oyama, and M. Shah, "Abnormal crowd behavior detection using social force model," in *CVPR*, 2009.

[5] C. Lu, J. Shi, and J. Jia, "Abnormal event detection at 150 fps in matlab," in *ICCV*, 2013.

[6] V. Saligrama and Z. Chen, "Video anomaly detection based on local statistical aggregates," in *CVPR*, 2012.

[7] Y. Cong, J. Yuan, and J. Liu, "Sparse reconstruction cost for abnormal event detection," in *CVPR*, 2011.

[8] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012.

[9] R. B. Girshick, "Fast R-CNN," in *ICCV*, 2015.

[10] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell, "Decaf: A deep convolutional activation feature for generic visual recognition," in *ICML*, 2014.

[11] A. Sharif Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "CNN features off-the-shelf: An astounding baseline for recognition," in *CVPRW*, 2014.

[12] B. Zhou, À. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning deep features for scene recognition using Places Database," in *NIPS*, 2014.

[13] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition," in *NIPS*, 2014.

[14] H. Mousavi, S. Mohammadi, A. Perina, R. Chellali, and V. Murino, "Analyzing tracklets for the detection of abnormal crowd behavior," in *WACV*, 2015.

[15] M. Ravanbakhsh, M. Nabi, H. Mousavi, E. Sangineto, and N. Sebe, "Plug-and-play cnn for crowd motion analysis: An application in abnormal event detection," *arXiv:1610.00307*, 2016.

[16] M. Sabokrou, M. Fayyaz, M. Fathy, and R. Klette, "Fully convolutional neural network for fast anomaly detection in crowded scenes," *arXiv:1609.00866*, 2016.

[17] D. Xu, Y. Yan, E. Ricci, and N. Sebe, "Detecting anomalous events in videos by learning deep representations of appearance and motion," *CVIU*, 2016.

[18] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio, "Generative adversarial nets," in *NIPS*, 2014.

[19] R. Raghavendra, M. Cristani, A. Del Bue, E. Sangineto, and V. Murino, "Anomaly detection in crowded scenes: A novel framework based on swarm optimization and social force modeling," in *Modeling, Simulation and Visual Analysis of Crowds*. 2013.

[20] H. Rabiee et al, "Novel dataset for fine-grained abnormal behavior understanding in crowd," in *AVSS*, 2016.

[21] H. Mousavi, M. Nabi, H. K. Galoogahi, A. Perina, and V. Murino, "Abnormality detection with improved histogram of oriented tracklets," in *ICIAP*, 2015.

[22] H. Rabiee et al, "Crowd behavior representation: an attribute-based approach," *SpringerPlus*, vol. 5, no. 1, pp. 1179, 2016.

[23] H. Rabiee et al, "Detection and localization of crowd behavior using a novel tracklet-based model," *Journal of Machine Learning and Cybernetics (JMLC)*, 2017.

[24] X. Huang, W. Wang, G. Shen, X. Feng, and X. Kong, "Crowd activity classification using category constrained correlated topic model," *KSII TIIS*, 2016.

[25] Tim Salimans, I. J. Goodfellow, Wo. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training GANs," in *NIPS*, 2016.

[26] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *ICLR*, 2016.

[27] P. Isola, J. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *CVPR*, 2017.

[28] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert, "High accuracy optical flow estimation based on a theory for warping," in *ECCV*, 2004.

[29] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, et al., "Imagenet large scale visual recognition challenge," *IJCV*, 2015.

[30] H. Mousavi, M. Nabi, H. Kiani, A. Perina, and V. Murino, "Crowd motion monitoring using tracklet-based commotion measure," in *ICIP*, 2015.

[31] H. Bilen, B. Fernando, E. Gavves, A. Vedaldi, and S. Gould, "Dynamic image networks for action recognition," in *CVPR*, 2016.