Learning Deep Representations of Appearance and Motion for Anomalous Event Detection

Dan Xu[1], Elisa Ricci[2]
danxuhk@gmail.com,eliricci@fbk.eu
Yan Yan[1,3], Jingkuan Song[1], Nicu Sebe[1]
yan@disi.unitn.it,jingkuan.song@unitn.it,sebe@disi.unitn.it

[1] DISI, University of Trento, Italy
[2] Fondazione Bruno Kessler, Trento, Italy
[3] Advanced Digital Sciences Center, UIUC Singapore, Singapore

Introduction. A fundamental challenge in intelligent video surveillance is to automatically detect abnormal events in long video streams. This problem has attracted considerable attentions in recent years. In this paper we propose a novel Appearance and Motion DeepNet (AMDN) framework for discovering anomalous activities in complex video surveillance scenes. Opposite to previous works [1, 2], instead of using hand-crafted features to model activity patterns, we propose to learn discriminative feature representations of both appearance and motion patterns in a fully unsupervised manner. A novel approach based on stacked denoising autoencoders (SDAE) [3] is introduced to achieve this goal.

Contributions. i) As far as we know, we are the first to introduce an unsupervised deep learning framework to automatically construct discriminative representations for video anomaly detection. ii) We propose a new approach to learn appearance and motion features as well as their correlations. Deep learning methods for combining multiple modalities have been investigated in previous works. However, to our knowledge, this is the first work where multimodal deep learning is applied to anomalous event detection. iii) A double fusion scheme is proposed to combine appearance and motion features for discovering unusual activities. iv) Our method is validated on challenging anomaly detection datasets and we obtain very competitive performance compared with the state-of-the-art.

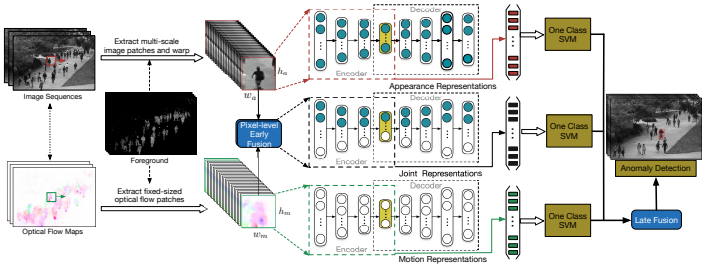


Figure 1: Overview of the proposed AMDN method for anomalous event detection. The proposed AMDN structure consists of three SDAE pipelines corresponding to different types of low-level inputs.

AMDN for Abnormal Event Detection. An overview of the proposed AMDN is shown in Fig. 1. Low-level visual information including still image patches and dynamic motion fields represented with optical flow is used as input of two separate networks, to first learn appearance and motion features, respectively. To further investigate the correlations between appearance and motion, early fusion is performed by combining image pixels with their corresponding optical flow to learn a joint representation. Finally, for abnormal event prediction, a late fusion strategy is introduced to combine the anomaly scores predicted by multiple one-class SVM classifiers, each corresponding to one of the three learned feature representations.

To learn the feature representations, we use three SDAE pipelines corresponding to different types of low-level inputs. The three SDAE networks learn appearance and motion features as well as a joint representation of them. We show the basic structures of the proposed SDAE networks in Fig. 2 (a) and (b). Each SDAE consists of two parts: encoder and decoder. For the encoder part, we use an over-complete set of filters in the first layer to capture a representative information from the data. Then, the number of neurons is reduced by half in the next layer until reaching the "bottleneck" hidden layer. The decoder part has a symmetric structure with respect to the encoder part.

We train the AMDN with two steps: pretraining and fine-tuning. The layer-wise pretraining learns one single denoising auto-encoder at a time with sparsity constraints. The input is corrupted to learn the mapping function, which is then used to produce the representation for the next layer with uncorrupted inputs. By using a greedy layer-wise pretraining, the denoising autoencoders can be stacked to build a multi-layer feedforward deep neural network, *i.e.* a stacked denoising autoencoder. Then fine-tuning is used to adjust parameters over the whole network.

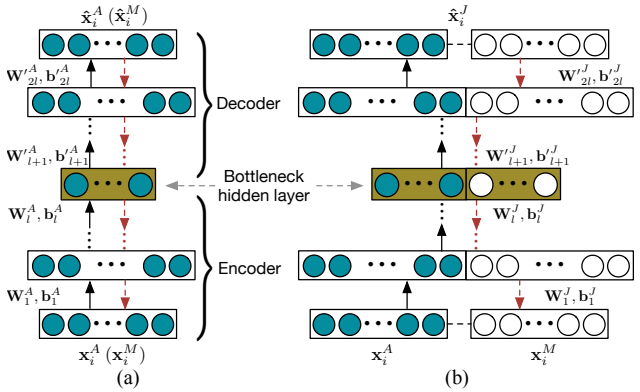
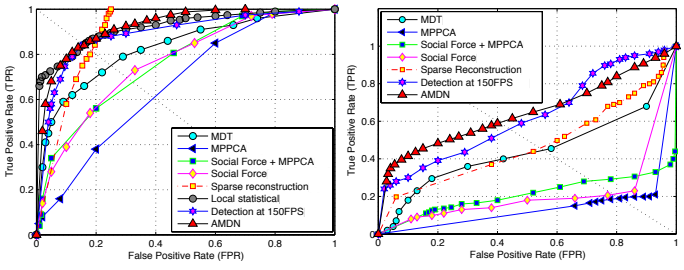


Figure 2: The structure of (a) the appearance and motion, and (b) the joint representation learning pipelines.

We formulate the video anomaly detection problem as a patch-based binary categorization problem, *i.e.* given a test frame we obtain $M_I \times N_I$ patches via sliding window with a stride d and classify each patch as corresponding to a normal or abnormal region. Specifically, given each test patch t we compute three anomaly scores $A^k(\mathbf{s}_t^k)$, $k \in \{A,M,J\}$, using one-class SVM models and the features representations $\mathbf{s}_t^k$ computed with the SDAEs. The three scores are then linearly combined to obtain the final anomaly score $\mathcal{A}(\mathbf{s}_t^k) = \sum_{k\in\{A,M,J\}} \alpha^k A^k(\mathbf{s}_t^k)$ ($k \in \{A,M,J\}$ corresponds to appearance, motion and joint representation, respectively). The weight vector α^k is automatically learned via an unsupervised late fusion scheme. Then for each patch t, we identify if it corresponds to an abnormal activity by computing the associated anomaly score $\mathcal{A}(\mathbf{s}_t^k)$ and comparing it with a threshold η, *i.e.* $\mathcal{A}(\mathbf{s}_t^k) \underset{abnormal}{\overset{normal}{\lessgtr}} \eta$.



(a) Frame-level ROC curve of PED1 Dataset (b) Pixel-level ROC curve of PED1 Dataset

Figure 3: UCSD dataset (Ped1 sequence): comparison of frame-level and pixel-level anomaly detection results with state of the art methods.

Results. Two publicly available datasets, the UCSD (Ped1 and Ped2) dataset and the Train dataset are used to evaluate the performance of the proposed approach. Fig. 3 (a) and (b) show the frame-level and pixel-level detection results on Ped1. The ROC curve is produced by varying the threshold parameter η. It is evident that our method outperforms most previous methods and that its performance are very competitive with the best two baselines. The proposed method is also evaluated on the Train dataset showing promising results (see the main paper).

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

[2] V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos. Anomaly detection in crowded scenes. In *CVPR*, 2010.

[3] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P. Manzagol. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *JMLR*, 11:3371–3408, 2010.