Computer Vision and Image Understanding xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu

Deep-anomaly: Fully convolutional neural network for fast anomaly detection in crowded scenes

Mohammad Sabokrou*,[1],[a], Mohsen Fayyaz[1],[b], Mahmood Fathy[a], Zahra. Moayed[c], Reinhard Klette[c]

[a] School of Computer Science, Institute for Research in Fundamental Sciences (IPM), Tehran PO Box 19395-5746, Iran
[b] University of Bonn, Germany
[c] School of Engineering, Computer and Mathematical Sciences, EEE Department, Auckland University of Technology, Auckland, New Zealand

ARTICLE INFO

Keywords:
Video anomaly detection
CNN
Transfer learning
Real-time processing

ABSTRACT

The detection of abnormal behaviour in crowded scenes has to deal with many challenges. This paper presents an efficient method for detection and localization of anomalies in videos. Using *fully convolutional neural networks* (FCNs) and temporal data, a pre-trained supervised FCN is transferred into an unsupervised FCN ensuring the detection of (global) anomalies in scenes. High performance in terms of speed and accuracy is achieved by investigating the cascaded detection as a result of reducing computation complexities. This FCN-based architecture addresses two main tasks, feature representation and cascaded outlier detection. Experimental results on two benchmarks suggest that the proposed method outperforms existing methods in terms of accuracy regarding detection and localization.

1. Introduction

The use of surveillance cameras requires that computer vision technologies need to be involved in the analysis of very large volumes of video data. The detection of anomalies in captured scenes is one of the applications in this area.

Anomaly detection and localization is a challenging task in video analysis already due to the fact that the definition of “anomaly” is subjective, or context-dependent. In general, an event is considered to identify an “anomaly” when it occurs rarely, or unexpected; for example, see Sabokrou et al. (2017b).

Compared to the previously published deep-cascade method in Sabokrou et al. (2017b), this paper proposes and evaluates a different and new method for anomaly detection. Here we introduce and study a modified pre-trained *convolutional neural network* (CNN) for detecting and localizing anomalies. In difference to Sabokrou et al. (2017b), the considered CNN is not trained from scratch but “just” fine-tuned. More in detail, for processing a video frame, Sabokrou et al. (2017b) outlined a method where the frame was first divided into a set of patches, then the anomaly detection was organised based on levels of patches. In difference to that, the input of the proposed CNN algorithm is a full video frame in this paper. As a brief preview, the new method is methodically simpler but faster in both the training and testing phase where the accuracy of anomaly detection is comparable to the accuracy of the method presented in Sabokrou et al. (2017b).

In the context of crowd scene videos, anomalies are formed by rare shapes or rare motions. Due to the fact that looking for unknown shapes or motions is a time-consuming task, state-of-the-art approaches learn regions or patches of normal frames as reference models. Indeed, these reference models include normal motion or shapes of every region of the training data. In the testing phase, those regions which differ from the *normal model* are considered to be abnormal. Classifying these regions into normal and abnormal requires extensive sets of training samples in order to describe the properties of each region efficiently.

There are numerous ways to describe region properties. Trajectory-based methods have been used to define behaviours of objects. Recently, for modeling spatio-temporal properties of video data, low-level features such as the *histogram of gradients* (HoG) or the *histogram of optic flows* (HoF) are used. These trajectory-based methods have two main disadvantages. They cannot handle occlusion problems, and they also suffer from high complexity, especially in crowded scenes.

CNNs proved recently to be useful for defining effective data analysis techniques for various applications. CNN-based approaches outperformed state-of-the-art methods in different areas including image classification (Krizhevsky et al., 2012), object detection (Girshick et al., 2014), or activity recognition (Simonyan and Zisserman, 0000). It is

* Corresponding author.
E-mail addresses: sabokro@ipm.ir (M. Sabokrou), fayyaz@iai.uni-bonn.de (M. Fayyaz), mfathy@ipm.ir (M. Fathy), zmoayed@aut.ac.nz (Z. Moayed), rklette@aut.ac.nz (R. Klette).
[1] Equal Collaboration.

https://doi.org/10.1016/j.cviu.2018.02.006
Received 3 October 2017; Received in revised form 16 January 2018; Accepted 15 February 2018
1077-3142/ © 2018 Elsevier Inc. All rights reserved.

ARTICLE IN PRESS

argued that handcrafted features cannot efficiently represent normal videos (Sabokrou et al., 2016; 2015; 2017a; Xu et al., 2015). In spite of these benefits, CNNs are computationally slow, especially when considering block-wise methods (Girshick et al., 2014; Giusti et al., 2013). Thus, dividing a video into a set of patches and representing them by using CNNs, should be followed by a further analysis with taking care about possible ways of speed-ups.

Major problems in anomaly detection using CNNs are as follows:

1. Too slow for patch-based methods; thus, CNN is considered as being a time-consuming procedure.
2. Training a CNN is totally supervised learning; thus, the detection of anomalies in real-world videos suffers from a basic impossibility of training large sets of samples from non-existing classes of anomalies.

Due to these difficulties, there is a recent trend to optimize CNN-based algorithms in order to be applicable in practice. Faster-RCNN (Ren et al., 2015) takes advantage of convolutional layers to have a feature map of every region in the input data, in order to detect the objects. For semantic segmentation, methods such as Shelhamer et al. (2016), Long et al. (2015) use *fully convolutional networks* (FCNs) for traditional CNNs to extract regional features. Making traditional classification CNNs to work as a fully convolutional network and using a regional feature extractor reduces computation costs. In general, as CNNs or FCNs are supervised methods, neither CNNs nor FCNs are capable for solving anomaly detection tasks,

To overcome aforementioned problems, we propose a new FCN-based structure to extract distinctive features of video regions. This new approach includes several initial convolutional layers of a pre-trained CNN using an AlexNet model (Krizhevsky et al., 2012) and an additional convolutional layer. AlexNet, similar to Zhou et al. (2014), is a pre-trained model proposed for image classification by using ImageNet (Deng et al., 2009; ImageNet, 2017) and the MIT places dataset (MIT places database, 2017). Extracted features, by following this approach, are sufficiently discriminative for anomaly detection in video data.

In general, entire frames are fed to the proposed FCN. As a result, features of all regions are extracted efficiently. By analysing the output, anomalies in the video are extracted and localized. The processes of convolution and pooling, in all of the CNN layers, run concurrently. A standard NVIDIA TITAN GPU processes ≈ 370 *frames per second* (fps) when analyzing (low-resolution) frames of size 320×240. This is considered to be "very fast".

The main contributions of this paper are as follows:

- To the best of our knowledge, this paper is one of the first where FCN is used for anomaly detection.
- We adapt a pre-trained classification CNN to an FCN for generating video regions to describe motion and shape concurrently.
- We propose a new FCN architecture for time-efficient anomaly detection and localization.
- The proposed method performs as well as state-of-the-art methods, but our method outperforms those with respect to time; we ensure real-time for typical applications.
- We achieved a processing speed of 370 fps on a standard GPU; this is about three times faster than the fastest existing method reported so far.

Section 2 provides a brief survey on existing work. We present the proposed method in Section 3 including the overall scheme of our method, and also details for anomaly detection and localization, and for the evaluation of different layers of the CNN for performance optimization. Qualitative and quantitative experiments are described in Section 4. Section 5 concludes.

2. Related work

Object trajectory estimation is often of interest in cases of anomaly detection; see Jiang et al. (2011), Wu et al. (2010), Piciarelli and Foresti (2006), Piciarelli et al. (2008), Antonakaki et al. (2009), Calderara et al. (2011), Morris and Trivedi (2011), Hu et al. (2006) and Tung et al. (2011). An object shows an anomaly if it does not follow learned normal trajectories. This approach usually suffers from many weaknesses, such as disability to efficiently handle occlusions, and being too complex for processing crowded scenes.

To avoid these two weaknesses, it is proposed to use spatio-temporal low level features such as optical flow or gradients. Zhang et al. (2005) use a *Markov random field* (MRF) to model the normal patterns of a video with respect to a number of features, such as rarity, unexpectedness, and relevance. Boiman and Irani (2007) consider an event as being abnormal if its reconstruction is impossible by using previous observations only. Adam et al. (2008) use an exponential distribution for modeling the histograms of optical flow in local regions.

A *mixture of dynamic textures* (MDT) is proposed by Mahadevan et al. (2010) for representing a video. In this method, the represented features fit into a Gaussian mixture model. In Li et al. (2014), the MDT is extended and explained in more details. Kim and Grauman (2009) exploit a *mixture of probabilistic PCA* (MPPCA) model for representing local optical flow patterns. They also use an MRF for learning the normal patterns.

A method based on motion properties of pixels for behavior modeling is proposed by Benezeth et al. (2009). They described the video by learning a co-occurrence matrix for normal events across space-time. In Kratz and Nishino (2009), a Gaussian model is fitted into spatio-temporal gradient features, and a *hidden Markov model* (HMM) is used for detecting the abnormal events.

Mehran et al. (2009) introduce *social force* (SF) as an efficient technique for abnormal motion modeling of crowds. Detection of abnormal behavior, using a method based on spatial-temporal oriented energy filtering, is proposed by Zaharescu and Wildes (2010).

Cong et al. (2011) construct an over-complete normal basis set from normal data. A patch is considered to be abnormal if it is impossible reconstructing it with this basis set.

In Antic and Ommer (2011), a scene parsing approach is proposed by Antic et al. All object hypotheses for the foreground of a frame are explained by normal training. Those hypotheses, that cannot be explained by normal training, are considered as showing an anomaly. Saligrama et al. propose in Saligrama and Chen (2012) a method based on clustering of test data using optic-flow features. Ullah and Conci (2012) introduced an approach based on a cut/max-flow algorithm for segmenting crowd motion. If a flow does not follow the regular motion model, it is considered as being an anomaly. Lu et al. (2013) propose a fast (140–150 fps) anomaly detection method based on sparse representation.

In Roshtkhari and Levine (2013a), an extension of the *bag of video words* (BOV) approach is used by Roshtkhari et al. A context-aware anomaly detection algorithm is proposed in Zhu et al. (2013) where authors represent a video using motions and the context of the video. In Cong et al. (2013), a method for modeling both motion and shape with respect to a descriptor (named "motion context") is proposed; authors consider anomaly detection as a matching problem. Roshtkhari and Levine (2013b) introduce a method for learning dominant events of a video by using the construction of a hierarchical codebook. Ullah et al. (2013) learn an MLP neural network using trained particles to extract the video behavior. A *Gaussian mixture model* (GMM) is exploited for learning the behavior of particles using extracted features. In addition, in Ullah et al. (2014b), an MLP neural network for extracting corner features from normal training samples is proposed; authors also label the test samples using that MLP.

Ullah et al. (2014a) extract corner features and analyze them based on their motion properties by an enthalpy model, using a random forest

دریافت فوری ← متن کامل مقاله

✓ امکان دانلود نسخه تمام متن مقالات انگلیسی
✓ امکان دانلود نسخه ترجمه شده مقالات
✓ پذیرش سفارش ترجمه تخصصی
✓ امکان جستجو در آرشیو جامعی از صدها موضوع و هزاران مقاله
✓ امکان دانلود رایگان ۲ صفحه اول هر مقاله
✓ امکان پرداخت اینترنتی با کلیه کارت های عضو شتاب
✓ دانلود فوری مقاله پس از پرداخت آنلاین
✓ پشتیبانی کامل خرید با بهره مندی از سیستم هوشمند رهگیری سفارشات

ISIArticles
مرجع مقالات تخصصی ایران