Neurocomputing 143 (2014) 144–152

Contents lists available at ScienceDirect

Neurocomputing

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

Video anomaly detection based on a hierarchical activity discovery within spatio-temporal contexts

Dan Xu [a,b,c], Rui Song [d], Xinyu Wu [a,b,*], Nannan Li [a], Wei Feng [a], Huihuan Qian [b,c]

[a] Guangdong Provincial Key Lab. of Robotics and Intelligent System, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, China
[b] Department of Mechanical and Automation Engineering, The Chinese University of Hong Kong, Hong Kong, China
[c] Center for Research on Robotics and Smart City, The Chinese University of Hong Kong & Smart China, Hong Kong, China
[d] School of Control Science and Engineering, Shandong University, China

ARTICLE INFO

Article history:
Received 8 November 2013
Received in revised form
12 May 2014
Accepted 3 June 2014
Communicated by D. Tao
Available online 18 June 2014

Keywords:
Visual surveillance
Video anomaly detection
Hierarchical discovery
Energy function

ABSTRACT

In this paper, we present a novel approach for video-anomaly detection in crowded and complicated scenes. The proposed approach detects anomalies based on a hierarchical activity-pattern discovery framework, comprehensively considering both global and local spatio-temporal contexts. The discovery is a coarse-to-fine learning process with unsupervised methods for automatically constructing normal activity patterns at different levels. A unified anomaly energy function is designed based on these discovered activity patterns to identify the abnormal level of an input motion pattern. We demonstrate the effectiveness of the proposed method on the UCSD anomaly-detection datasets and compare the performance with existing work.

1. Introduction

Video anomaly detection has become an important research aspect in the area of intelligent visual surveillance due to growing security needs. With the application of video surveillance in modern life growing gradually [1,2], drawbacks of conventional surveillance are revealing themselves; for example, spending a long time staring at monitors causes operators fatigue and inattention, sometimes causing them to neglect certain underlying dangerous occurrences. Additionally, since existing surveillance functions tend to capture evidence in a surveying manner, they cannot provide a warning when risk events are forming. Although automated anomaly-detection has attractive potential, it is also one of difficult problems in video analysis. Firstly, unusual events are rare, difficult to describe, and often subtle; secondly, visual behavior is diverse and complex in a realistic and unconstrained environment; thirdly, the description and definition of normality and abnormality, have high uncertainty and depend on changing visual contexts [3]. Many researchers have been focusing on this area in recent years, and provide various possible solutions for the problem, including unsupervised approaches based on modeling from underlying visual features [4,5], and supervised approaches based on training models or templates from tagged behaviors [6–8]. However, the uncertainty of abnormal activity description and scene-complexity makes anomaly detection a challenging problem.

For detecting video anomalies, one category of popular approaches in the literature is the tracking-based methods [9–11]. The main idea of such methods is to first analyze and model normal trajectories collected by tracking individual moving-objects in the video, and then to detect anomalous object-motions whose trajectories are deviating from the normal model. These methods can obtain promising results in less-cluttered scenes, with only a few people; however, in dense crowds, achieving robust tracking is a quite difficult task because of serious occlusion problems, which heavily degrade the performance of anomaly detection.

To avoid the aforementioned limitations, the other category of methods addresses the problem by learning activity patterns from low-level visual features, such as motion and texture. Zhang et al. propose a slow feature analysis (SFA) based on a feature-learning method which can extract useful motion patterns from the video sequences for activity description [1]. Andrade et al. model crowd scenes with Hidden Markov Models combined with spectral clustering for detecting unusual events [12]. Mehran et al. propose

* Corresponding author at: Guangdong Provincial Key Lab. of Robotics and Intelligent System, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, 1068 Xueyuan Avenue, Shenzhen University Town, Shenzhen, P.R. China. Tel.: +86 8639 2135; fax: +86 8639 2194.
E-mail addresses: danxuhk@gmail.com (D. Xu), rsong@sdu.edu.cn (R. Song), xy.wu@siat.ac.cn (X. Wu), nn.li@siat.ac.cn (N. Li), wei.feng@siat.ac.cn (W. Feng), hhqian@mae.cuhk.edu.hk (H. Qian).

http://dx.doi.org/10.1016/j.neucom.2014.06.011
0925-2312/

to model crowd activity patterns for anomaly detection by using a "social force" model based on optical flow feature representation [13]. Kim et al. use a mixed dynamic texture model to detect spatial and temporal anomalies through a joint modeling of the appearance and dynamics of the scene [14]. However, these methods model activity patterns only considering the local context or the global context, which leads to a lack of global information or local location relationships for the simultaneous perception of both local and global abnormal motion patterns.

In this paper, we aim to detect anomalies by comprehensively considering both global and local spatio-temporal contexts. A hierarchical framework of learning activity patterns is proposed to achieve this task. Under the global context, we discover atomic activity patterns from low-level optical flow features, and the distributions of the atomic activity patterns are modeled for higher-level activity representation. Then, salient activity patterns are discovered under the local context. The two layers of discovery both adopt unsupervised ways without any priori knowledge of the anomalies. Finally, we design a unified abnormal energy function to detect global and local pattern anomalies. The main contribution of this work can be summarized into three aspects:

- A new video anomaly detection framework is proposed using unsupervised learning techniques, which allows one to build a detection model without any hypothesis of anomalies and any specification of anomalous activity classes.
- Both a top-down and a bottom-up approach are utilized in the process of hierarchical discovery of activities, which makes the system can learn multi-level activity patterns for the perception of global, local, and co-occurrence anomalous events in complicated scenes.
- The discovered patterns are directly used to construct energy function for detection, without extra complex calculation steps, which brings about low calculation complexity in detection, and makes the system perform in a near real-time speed in a standard PC.

The rest of this paper is organized as follows. Section 2 gives a brief introduction of related work on video-anomaly detection in the literature. Section 3 provides an overview of the proposed method. The problem of feature representation is addressed in Section 3.1. Section 3.2 describes the process of hierarchical activity discovery, which includes a discovery of atomic activities in a global context and a discovery of salient activities in the local context. Section 3.3 illustrates the details of the unified energy function for anomaly detection. The performance evaluation for the proposed method is presented in Section 4, and we finally conclude the paper in Section 5.

2. Related work

Many researchers have been focusing on video-anomaly detection in the past decade, and a comprehensive survey of this problem can be found in several review papers [15,16]. The existing techniques of anomaly detection in the literature can be generally divided into two broad categories: supervised and unsupervised approaches. We briefly describe the related work from these two aspects.

Supervised approaches usually first train models or templates from tagged behaviors or objects, and then perform detection in test samples with the learned models or templates. Such methods build the model depending on pre-defined behavior classes containing both normal and anomalous ones, which can detect specific anomalies under rigorous restrictions of video scene conditions [6,17]. Raz et al. propose a state transition model to perform sematic anomaly detection in dynamical data feeds from online data sources [18]. However, in real-life scenarios, this is inferior for the anomaly detection problem due to the lack of a priori knowledge about types of abnormal activities. In addition, the training video data with clear activity labels for supervised model construction is also difficult to acquire.

The unsupervised methods are usually based on statistical learning. Various modern statistical machine learning techniques have been used to solve computer vision problems such as image classification [19,20], scene understanding [21], and cartoon animation [22,23]. Nonnegative matrix factorization (NMF) [24–26] is well studied for high-dimensional data representation, which has also been introduced by researchers for anomaly detection in recent years [27]. For visual anomaly modeling and detection, a number of techniques have also been proposed using unsupervised learning, which can be further categorized into two different types, according to what kinds of behavior representations are adopted. One of them is based on trajectory modeling. The main idea of such methods is analyzing and modeling normal trajectories obtained by tracking individual objects in training videos, then declaring behaviors to be abnormal when they deviate from frequent tracks [28–31]. These methods can obtain promising results when foreground objects are easy to detect and trace, especially in indoor environments; however, they suffer from occlusion problems which occur commonly in densely crowded scenes, for which detection results are not very promising.

To address the limitations encountered by object-tracking-based methods, some authors have proposed various learning methods based on underlying visual characteristics other than motion trajectories, such as dense optical flow [32] and spatial–temporal gradients [14,13,33]. Louis et al. first represent localized motion patterns with 3D Gaussian distributions of spatial–temporal gradients, then use a coupled HMM to describe temporal and spatial relationships between local motion patterns [34]. Mehran et al. propose to characterize crowd behavior using concepts, such as social force, which are combined with a latent Dirichlet allocation (LDA) model for anomaly detection [35]. Kim et al. use a mixture of probabilistic principle components to represent typical local patterns, which are enforced with a space–time Markov random field (MRF) to detect abnormal behaviors locally and globally. Further, their model can update incrementally to adapt to environmental change [14]. All of these approaches mentioned above focus uniquely on motion information, ignoring abnormality information, due to variations in object appearance. Xiong et al. build an image potential-energy model based on the analysis of the apparent size of objects, which is used to detect two types of crowd anomaly events, such as gathering and running [36]. Mahadevan et al. propose to jointly model the appearance and dynamics of crowded scenes, using mixtures of dynamic textures (MDT), which explicitly investigate both temporal and spatial anomalies [35]; the main drawback of this method, however, is heavy computation. To take advantage of the complementary effect of different feature representations, Reddy et al. model crowded scenes by fusing three types of features, including speed, size, and texture for anomaly detection [37].

The proposed work can be classified into the class of unsupervised approaches. Similar to the method [38] which models two levels of video activities, including atomic activities and interactions, we propose a hierarchical framework for activity discovery in the video. The process of hierarchical discovery uses both a top-down and a bottom-up approach, which is different from the top-down method used by Wang et al. [38]. In addition, our paper considers both global and local spatio-temporal contexts in the discovery process. Thus, our anomaly-detection model, which is built from multi-level activity patterns learned from global and local contexts, can detect global, local, and co-occurrence anomalous events in crowded and complicated scenes.