Pattern Recognition 73 (2018) 99–110

Contents lists available at ScienceDirect

Pattern Recognition

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

Structured dictionary learning for abnormal event detection in crowded scenes

Yuan Yuan, Yachuang Feng, Xiaoqiang Lu*

Center for OPTical IMagery Analysis and Learning (OPTIMAL), State Key Laboratory of Transient Optics and Photonics, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an 710119, Shaanxi, PR China

ARTICLE INFO

Article history:
Received 19 November 2015
Revised 2 May 2017
Accepted 2 August 2017
Available online 3 August 2017

Keywords:
Video surveillance
Abnormal event detection
Dictionary learning
Sparse representation
Reference event

ABSTRACT

Abnormal event detection is now a widely concerned research topic, especially for crowded scenes. In recent years, many dictionary learning algorithms have been developed to learn normal event regularities, and have presented promising performance for abnormal event detection. However, they seldom consider the structural information, which plays important roles in many computer vision tasks, such as image denoising and segmentation. In this paper, structural information is explored within a sparse representation framework. On the one hand, we introduce a new concept named reference event, which indicates the potential event patterns in normal video events. Compared with abnormal events, normal ones are more likely to approximate these reference events. On the other hand, a smoothness regularization is constructed to describe the relationships among video events. The relationships consist of both similarities in the feature space and relative positions in the video sequences. In this case, video events related to each other are more likely to possess similar representations. The structured dictionary and sparse representation coefficients are optimized through an iterative updating strategy. In the testing phase, abnormal events are identified as samples which cannot be well represented using the learned dictionary. Extensive experiments and comparisons with state-of-the-art algorithms have been conducted to prove the effectiveness of the proposed algorithm.

1. Introduction

Nowadays, video surveillance equipments have been widely used due to the heightened security concerns. However, manual analysis for the increasing volume of video dataset is costly. Meanwhile, most manual efforts are unnecessary (over 99.9% [1]), because the majority of video events are normal. Consequently, it is vital to detect abnormal events automatically. Generally, abnormal events can be divided into two categories: uncrowded and crowded anomalies. By extracting object trajectories, uncrowded abnormal events can be accurately detected. Along with the increase of crowd density, the analysis of individual moving objects is impractical due to serious clutter and occlusion. As a severe challenge, abnormal event detection in crowded scenes draws more and more research interests.

In the field of anomaly detection, it is hard to design a general framework for detecting abnormal events in diverse situations, since anomalies in one scenario may be normal in another. Meanwhile, it is scarcely possible to list all abnormal patterns in a given scene. These properties are much more obvious in crowded scenes, since behaviors become much more complicated. For these reasons, one common solution is to learn normal patterns from training videos which do not contain any abnormal events, and detect abnormal ones by finding dissimilarities.

Latest developments in abnormal event detection mainly focus on two aspects: event representation and model learning. *Event representation* aims to extract distinguishable features for different event categories. One intuitive event representation is object trajectory [2–6] which benefits from the high-level semantic. However, trajectory based features are no longer effective in crowded scenes, since tracking is unreliable with inevitable overlap and occlusion. Alternatively, a number of cuboid based features have been proposed, such as saliency features [7], 3D video patches of temporal derivatives [8], spatio-temporal gradient [9], and chaotic invariant features [10]. Moreover, Mehran et al. [11] propose a *social force* (SF) model to analyze crowd behaviors. Adam et al. [12] adopt histograms to model optical flow probabilities at a group of fixed spatial locations. Cong et al. [13,14] describe video events with multi-scale histograms of optical flow. Attributed to the high crowd density, cuboid based features hold the major part in crowd abnormal event detection. *Model learning* is to learn normal event

* Corresponding author:
E-mail address: luxq666666@gmail.com (X. Lu).

http://dx.doi.org/10.1016/j.patcog.2017.08.001
0031-3203/

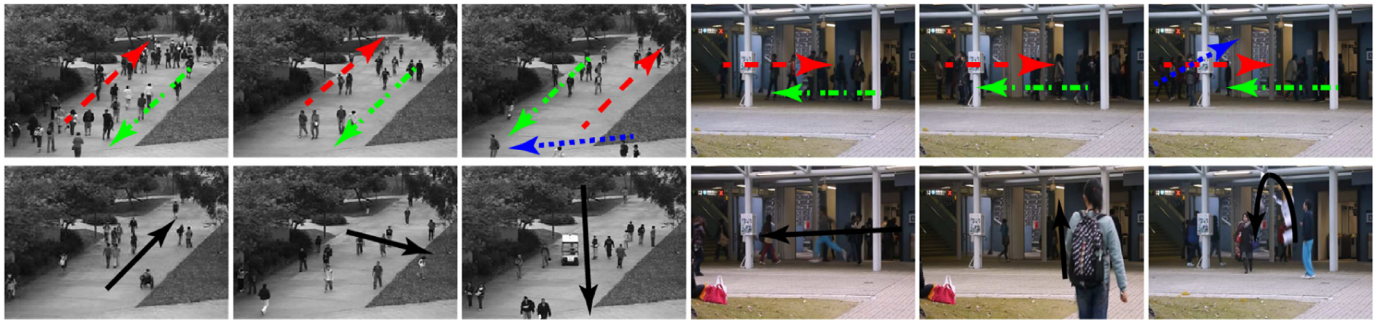

Fig. 1. Examples of normal and abnormal events. The top row shows normal events, and the bottom row displays abnormal events. Color dotted lines represent normal event patterns in the scenes, while black solid lines stand for abnormal event patterns. It's easy to find that normal events occur frequently in a set of regular patterns, while abnormal events do not share this commonality.

patterns from the training dataset, such as Laplacian eigenmap [15], Bayesian model [7], and infinite *hidden Markov model* (HMM) [16]. In [17], Kim et al. use a *mixture of probabilistic principal component analyzers* (MPPCA) to model local optical flow, and enforce the consistency by a spatio-temporal *Markov random field* (MRF) model. Vaswani et al. [18] adopt an HMM to model shape activities of objects. To analyze SF features, Mehran et al. [11] make use of a *latent Dirichlet allocation* (LDA) model [19,20]. In [21,22], a *mixture of dynamic textures* (MDT) is used to jointly model dynamics and appearances of crowded scenes. Although these methods achieve state-of-the-art performance, they always suffer from very high computational costs.

Recently, many algorithms based on *sparse representation* (SR) [13,23,24] have been proposed to deal with the abnormal event detection problems. The basic idea is that normal video events can be represented as sparse linear combinations of a set of bases learned from training videos. For example, Cong et al. [13] assume the bases used more frequently are more related to normal events. In [1], Lu et al. propose an efficient sparse combination framework, which achieves a very high detection rate. Mo et al. [25] propose a joint sparsity model, which deals with joint abnormal events involving multiple objects. In these algorithms, whether video events are abnormal is based on scores of the cost function. Specifically, normal events have small scores under the sparsity constraint, while abnormal events possess large ones.

SR based algorithms have been proven to be effective and efficient for abnormal event detection [1,13]. Nevertheless, they seldom consider the potential structural information of video events. In many computer vision tasks, structural information plays an important role and is widely used, such as denoising [26], segmentation [27], and hyperspectral unmixing [28]. For these reasons, we propose an algorithm which considers structural information in an SR framework. In this paper, the exploration of structural information mainly focuses on two aspects: the extraction of normal event patterns and the exploitation of data distribution. In a specific scenario, the patterns of abnormal events can be hardly predicted, but normal event patterns could be observed easily. Some examples of normal and abnormal events are shown in Fig. 1. We can find that normal events occur frequently in a set of regular patterns, while abnormal events are generally ruleless. In this paper, the regularly occurring normal event patterns are named as reference events. Video events dissimilar with these reference events are more likely to be abnormal. Meanwhile, data distributions are made use of for the purpose of anomaly detection. On the one hand, video events with similar features should have similar sparse representation coefficients. On the other hand, video events at adjacent spatio-temporal locations ought to possess similar representations. In this case, we develop a smoothness regularization to preserve the data distribution information in the sparse representation coefficients.

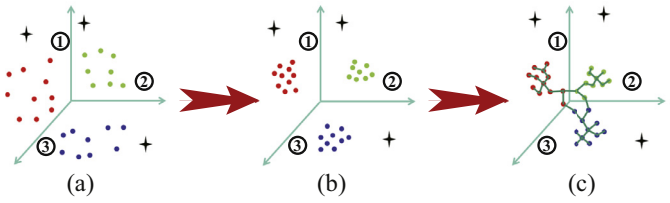

Fig. 2. Distribution of sparse representation coefficients under different conditions. Points with different colors and shapes are coefficient vectors of various video events, displaying in the coordinate space. There are three kinds of normal event patterns, which are represented by circles with different colors. Black stars stand for abnormal events. (a) The original sparse representation. (b) Sparse representation combined with reference events. (c) Sparse representation combined with both reference events and the smoothness regularization.

In this paper, we propose a new sparse representation framework, which combines both reference events and the smoothness regularization. As a result, sparse representation coefficients tend to present similar structure with the original video events. A simple illustration is shown in Fig. 2. In Fig. 2(a), coefficient vectors of the original SR distribute randomly in the coordinate space. In Fig. 2(b), coefficient vectors of normal events exhibit potential event patterns due to the introduction of reference events. With the smoothness regularization, data distribution of the original video events gets better preserved in Fig. 2(c). Meanwhile, we can find that the identification of abnormal events becomes easier.

Compared with state-of-the-art algorithms, contributions of this paper are listed as follows:

- The concept of reference event is introduced to represent major event patterns in normal videos. Generally, normal events have similar event patterns with these reference events, but abnormal ones present divergent patterns.
- A smoothness regularization is constructed to preserve the data distribution of video events. This regularization is inspired by the fact that similar video events at neighboring positions are more likely to have similar representation coefficients.
- A general SR framework is proposed to combine these reference events and the smoothness regularization. As a consequence, normal events have similar representations with their corresponding reference events, as well as their neighbors in both feature space and video sequences. Comparison results demonstrate that the proposed algorithm outperforms others.

The rest of this paper is organized as follows. Section 2 explains details of the proposed algorithm, including model learning and abnormal event detection. Comparison experiments with