Neurocomputing 204 (2016) 106–115

Contents lists available at ScienceDirect

Neurocomputing

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

Quaternion discrete cosine transformation signature analysis in crowd scenes for abnormal event detection

Huiwen Guo [a], Xinyu Wu [a,b,*], Shibo Cai [c], Nannan Li [a], Jun Cheng [a], Yen-Lun Chen [a]

[a] *Guangdong Provincial Key Lab of Robotics and Intelligent Systems, Shenzhen Institute of Advanced Technology, Chinese Academy of Science, China*
[b] *Department of Mechanical and Automation Engineering, The Chinese University of Hong Kong, Hong Kong, China*
[c] *Key Laboratory of E& M (Zhejiang University of Technology), Ministry of Education, China*

ARTICLE INFO

Article history:
Received 13 March 2015
Received in revised form
12 July 2015
Accepted 25 July 2015
Available online 8 April 2016

Keywords:
Video anomaly detection
Quaternion discrete cosine transformation
Spectral analysis
Multi-scale analysis

ABSTRACT

In this paper, an abnormal event detection approach inspired by the saliency attention mechanism of human visual system is presented. Conventionally, statistics-based methods suffer from visual scale, complexity of normal events and insufficiency of training data, for the reason that a normal behavior model established from normal video data is used to detect unusual behaviors with an assumption that anomalies are events with rare appearance. Instead, we make the assumption that anomalies are events that attract human attention. Temporal and spatial anomaly saliency are considered consistently by representing the pixel value in each frame as a quaternion, with weighted components that composed of intensity, contour, motion-speed and motion-direction feature. For each quaternion frame, Quaternion Discrete Cosine Transformation (QDCT) and signature operation are applied. The spatio-temporal anomaly saliency map is developed by inverse QDCT and Gaussian smoothing. By multi-scale analyzing, abnormal events appear at those areas with high saliency score. Experiments on typical datasets show that our method can achieve high accuracy results.

1. Introduction

Intelligent video surveillance has been a very important technique for monitoring dense crowd in public places for crowd disaster prevention, emergencies alarming, and the safety of life and property protection. This has drawn much attention of the researchers, for the benefit that not only speeding up the response time of security agencies, but also liberating a large number of security persons from tedious work of watching videos. However, abnormal event detection still faces problems. The basic problem is the universal definition of an abnormal event. On one hand, it is infeasible to provide the list of all abnormal events in a given scene due to the variety of abnormal events. On the other hand, some normal events are treated as abnormal in different scenes. An event is normal or not is closely related to the place, the time of the occurrence and the surrounding. Since the variety of surveillance place and abnormal events, we focus only on two typical types of abnormal events occured at crowd scenes, one is that wrong object movement at surveillance area, for example, cyclist moving at walking street, the other one is that wrong behaviors appearing on objects, such as escaping on square.

To detect these abnormal events, a popular idea is that these abnormal events are defined as the events occurring infrequently [1]. A probabilistic viewpoint could be adopted to an abnormal event, which is conformed by intuition. And normal event model is built underlying a statistics-based framework. Clustering-based method [2], reconstitution-based method [3] and inferring-based method [4] are frequently adopted. However, some big challenges are encountered by using these methods. Firstly, it makes anomalies dependent on the visual scale at which training data is defined. An abnormal behavior at fine visual scales may be perceived as normal when a larger scale is considered, and vice versa. Secondly, various normal event models are needed for different scenes. For instance, people appearing on a walking street is considered as a normal event, but people appearing on the highway is taken as an abnormal event. Thirdly, a large number of normal training datasets are needed to build normal-event models robustly.

However, the interesting thing is that a child without too much knowledge can find abnormal events in video only by a glance, such as gathering, vehicle on the sidewalk. These do not depend on the prior knowledge of the individual, simply because these events are very different from those around them. Although

[*] Corresponding author at: Guangdong Provincial Key Lab of Robotics and Intelligent Systems, Shenzhen Institute of Advanced Technology, Chinese Academy of Science, China.

E-mail addresses: hw.guo@siat.ac.cn (H. Guo), xy.wu@siat.ac.cn (X. Wu), ccc@zjut.edu.com (S. Cai), nn.li@siat.ac.cn (N. Li), jun.cheng@siat.ac.cn (J. Cheng), yl.chen@siat.ac.cn (Y.-L. Chen).

http://dx.doi.org/10.1016/j.neucom.2015.07.153
0925-2312/© 2016 Published by Elsevier B.V.

people are not deliberate, people neglect the brain's summary of normal events, and only care for special events. Thus, we hope utilize the inherent characteristics of the people and avoid the mentioned problems of traditional methods. Based on this, we define these abnormal events as events that attract attention in saliency view, which is inspired by the saliency attention mechanism of human visual system.

Primates have a remarkable ability to qualitatively interpret complex scenes or events in real time, despite the limited speed of the neuronal hardware available for such tasks [4]. Before further processing the huge amount of information, intermediate and higher visual processes appear to select a subset of the available sensory information [5], most likely to reduce the complexity of scene analysis [6]. The selected spatially circumscribed region is called the focus of attention [4], which scans the scene both in a rapid, bottom-up, saliency-driven, and task-independent manner as well as in a slower, top-down, volition-controlled, and task-dependent manner [6]. For the visual system in primates, visual input is first decomposed into a set of topographic feature maps. Competition within maps that are with different spatial location appeared, only those locations that locally stand out from their surround can persist [4]. In primates, such a map is believed to be located in the posterior parietal cortex [7] as well as in the various visual maps in the pulvinar nuclei of the thalamus [8]. It is plausible that the locations of abnormal events in surveillance videos are the same to the saliency attention region that located by human visual nervous system, according to the above research findings. The sparseness and rareness characteristic of the abnormal events result in the triumph from other normal events, which persists after intermediate and higher visual processes of human.

Anomaly in our definition is consistent with the popular definition to some extent that are based on comparison between events. Intuitively, our assumption can be demonstrated by some practical examples. For instance, some abnormal events, such as a rider on sidewalks, retrograding on the one-way street, car appearing on a sidewalk, or fighting on the street, attract our attention immediately.

In this work, we propose a saliency-attention-based approach to detect two kinds of abnormal events in crowd scenes, according to the definition that attractive events are anomalies. For the consistent consideration of temporal and spatial saliency attention in video, each frame is represented by a weighted quaternion image that have four channels, which are the frame intensity, object contour, motion speed and motion direction. By using Quaternion Discrete Cosine Transformation (QDCT) and signature operation for frequent components, new frequent components could be obtained. The spatio-temporal anomaly-saliency-attention map is obtained by the inverse QDCT operation and Gaussian smoothing with multi-scale analysis. The location of abnormal events appear at those areas with high anomaly saliency attention scores. The main contribution of this work can be summarized into the following aspects:

- A new assumption to abnormal events is introduced, which allows one to detect abnormal events in crowd scenes without any prior of anomalies and any specification of anomalous activity classes.
- By using quaternion frame, the low level feature representation of events consistent consideration of temporal and spatial information.
- There is no extra complex calculation steps, which result in low calculation complexity in detection, and make the system perform at 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 and spectral approaches. Section 3 provides a detailed description of the proposed method and gives some characteristics analysis. The performance evaluation for the proposed method is shown in Section 4, and conclusions and future work are given in Section 5.

2. Related work

In the past decade, many researchers have been focusing on abnormal events detection, and comprehensive surveys of this problem can be found in several review papers [9,10]. The existing approaches of abnormal events detection can be generally divided into two broad categories: supervised and unsupervised approaches, depending on the approaches applied for constructing the model.

Supervised-based methods usually establish the normal and abnormal behavior models from labeled video data, and can detect anomalous defined beforehand in the training phase. Information extraction [11], preprocess of features [12–14] and classifiers such as SVM [15] were usually used. However, the built models depend on pre-defined behavior classes containing both normal and anomalous ones, which can only detect specific anomalies under rigorous restrictions of video scene conditions [16,17]. State transition model has been used to perform sematic anomaly detection via online learning [18]. However, in real, complex environments [11,19] where it is impossible to specify the types of anomalies, they cannot work effectively.

Unsupervised-based methods only provide normal instances, the center idea of such approaches is to learn the normal behavior model either automatically or through a training process.

In the literature, one category of the existing approaches is based on trajectory, such as [20–25]. The significance of trajectory is clear, although the extraction process is complicated. By using object tracking method, typical extraction process obtain the spatio-temporal trajectories, which they are not robust in crowd scene. Wu [20] gets particle trajectories in crowd scene inspired by particle advection. The same procedure has been done in [21], which extracts part and short-time trajectories. To find the optimal spatio-temporal trajectories in 3D volume, [22] uses path optimization algorithm. As the trajectories are extracted, normal event model is obtained by clustering the trajectories. Ref. [23] presents a multi-level k-means clustering algorithm both in temporal and spatial aspects. Ref. [24] not only considers the dynamics of abnormal trajectories, but also considers the co-occurrence relationship between trajectories. Treat points in trajectories as graph node, spectral graph analysis is used to find the abnormal trajectories [25]. By whole training normal trajectories, [26] calculates the trajectory trending to predict unfinished trajectory online. The performance of the trajectory-based methods largely dependent on the extraction process of trajectories. In scenes with few people, these approaches can obtain precise detection results; however, in dense crowds, it is quite difficult to get robust result, although some improvement in articles has been proposed.

Another category of approaches utilizes statistics-learning based methods to build a normal behavior model for anomaly detection. Such methods rely on low-level features which are extracted from image patches or spatio-temporal video volumes (STVVs). These features include optical flow [2,27,28], histogram of optical flows [29], histogram of spatio-temporal gradients [30,31], texture of optical flow [32], and dynamic texture [28]. Furthermore, some scholars [33,34] integrate various features to obtain better detection results, taking advantage of mutual complements from the descriptions of motion related to different features. Compared to the aforementioned tracking-based methods, those approaches demonstrate their robustness in complex environments with dense crowds, since the low-level features which they