Expert Systems With Applications 54 (2016) 13–28

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Detecting rare events using Kullback–Leibler divergence: A weakly supervised approach

Jingxin Xu, Simon Denman*, Clinton Fookes, Sridha Sridharan

Queensland University of Technology, SAIVT Laboratory, Science and Engineering Faculty, 2 George Street, Brisbane, QLD 4000, Australia

ARTICLE INFO

Keywords:
Event detection
Weakly supervised learning
Kullback–Leibler divergence
Anomaly detection

ABSTRACT

Video surveillance infrastructure has been widely installed in public places for security purposes. However, live video feeds are typically monitored by human staff, making the detection of important events as they occur difficult. As such, an expert system that can automatically detect events of interest in surveillance footage is highly desirable. Although a number of approaches have been proposed, they have significant limitations: supervised approaches, which can detect a specific event, ideally require a large number of samples with the event spatially and temporally localised; while unsupervised approaches, which do not require this demanding annotation, can only detect whether an event is abnormal and not specific event types. To overcome these problems, we formulate a weakly-supervised approach using Kullback–Leibler (KL) divergence to detect rare events. The proposed approach leverages the sparse nature of the target events to its advantage, and we show that this data imbalance guarantees the existence of a decision boundary to separate samples that contain the target event from those that do not. This trait, combined with the coarse annotation used by weakly supervised learning (that only indicates approximately when an event occurs), greatly reduces the annotation burden while retaining the ability to detect specific events. Furthermore, the proposed classifier requires only a decision threshold, simplifying its use compared to other weakly supervised approaches. We show that the proposed approach outperforms state-of-the-art methods on a popular real-world traffic surveillance dataset, while preserving real time performance.

1. Introduction

Video surveillance infrastructure plays an important role in public security and safety, and as a result security cameras are installed in the majority of public places (e.g. airports, train stations, road ways, shopping malls, etc.). However, the detection of events of interest, or abnormal events, is still heavily reliant on human observers monitoring security footage. As such, the automatic detection of events from video surveillance has become an increasingly active area of research in recent years. However due to a number of challenges (e.g. occlusions, real-time requirements, low image resolution, ground-truth annotation, etc.), it remains an open problem.

Automatic event detection systems typically contain feature extraction and pattern classification components, although the boundary between these is becoming blurred in this deep learning age (Ji, Xu, Yang, & Yu, 2013a; Simonyan & Zisserman, 2014; Wang & Ji, 2015). With regards to feature extraction, approaches that depend on object tracking (Wang, Ma, Ng, & Grimson, 2011) have been eschewed in favour of approaches that extract features from local image regions to recognise actions and events (Adam, Rivlin, Shimshoni, & Reinitz, 2008; Mahadevan, Li, Bhalodia, & Vasconcelos, 2010; Nallaivarothayan, Fookes, Denman, & Sridharan, 2014; Roshtkhari & Levine, 2013; Wang, Ma, & Grimson, 2009; Xiang & Gong, 2006; Xu, Denman, Sridharan, Fookes, & Rana, 2011), as object tracking becomes increasingly difficult in crowded scenes with large numbers of occlusions.

The subsequent step of pattern classification, with which this paper is primarily concerned, can be broadly separated into two paradigms: unsupervised learning approaches, and supervised learning approaches. One-class unsupervised learning (Adam et al., 2008; Wang et al., 2009; Xu et al., 2011) has proven popular, and it essentially amounts to outlier detection. Such approaches assume that suspicious and/or emergency events occur at low frequencies, and enable systems to be trained with minimal data annotation, requiring footage that contains only normal events with no events explicitly labelled. However such systems are limited in that they cannot identify what type of event a detected event is; and finding

* Corresponding author. Tel.: +61 731389329.
E-mail addresses: j15.xu@qut.edu.au (J. Xu), s.denman@qut.edu.au (S. Denman), c.fookes@qut.edu.au (C. Fookes), s.sridharan@qut.edu.au (S. Sridharan).

http://dx.doi.org/10.1016/j.eswa.2016.01.035
0957-4174/

suitable footage for training that contains not only no abnormal events, but also instances of all "normal" events can be challenging, leading to events of little security interest being detected as abnormal.

To be able to detect and identify specific events, supervised learning approaches (Ji, Xu, Yang, & Yu, 2013b; Yuan, Liu, & Wu, 2009) are needed. Supervised approaches explicitly learn a model to detect an event of interest. However, the annotation required for supervised learning approaches is onerous. Typically, bounding boxes need to be annotated every frame for each activity, which is highly impractical in crowded and complex surveillance scenes. Furthermore, depending on the event in question, it may be difficult to obtain sufficient samples of the event for training.

Recently, (Hospedales, Li, Gong, & Xiang, 2011) introduced a third approach, "weakly supervised learning", with the proposal of a weakly supervised joint topic model (WSJTM). Weakly supervised learning is a special case of supervised learning where binary labels at a coarse level (in the case of Hospedales et al., 2011, annotation is performed by indicating whether a short video segments a few seconds in duration contain the target event or not) are used to indicate if an event of interest is present or not, without explicitly identifying exactly when or where the event occurs. Building on other approaches that use topic models (Hospedales, Gong, & Xiang, 2009; Wang et al., 2009), (Hospedales et al., 2011) used a discrete optical flow codebook to encode the events, such that a video clip is represented using a histogram of visual words. The approach is shown to be effective for detecting subtle events in traffic surveillance footages.

Following on from Hospedales et al. (2011), Xu, Denman, Sridharan, and Fookes (2015) proposed coupling a labelled Topic Model (Ramage, Hall, Nallapati, & Manning, 2009) with a feature descriptor that is specifically designed to exploit compressed data for highly efficient multi-person event detection. This approach was shown to offer much greater than real-time processing, while preserving state-of-the-art detection performance. Xu, Denman, Reddy, Fookes, and Sridharan (2013) also proposed a weakly supervised classifier constructed using random matrix theory for traffic event detection. However, these weakly supervised methods (Hospedales et al., 2011; Xu et al., 2013; Xu et al., 2015) all require a number of parameters to be set or initialised (for instance, both Hospedales et al., 2011 and Xu et al., 2015 are forms of topic models and require Dirichlet parameters to be set), making deployment problematic and resulting in unstable results for small training datasets.

In this paper, we propose a new weakly supervised learning approach that requires on only a single parameter, and leverages the frequent imbalance between positive and negative samples to our advantage. We propose the use of Kullback–Leibler (KL) divergence (Kullback & Leibler, 1951) to detect short video clips that contain the event of interest, and prove that through this imbalance of data (having many more negative samples than positive) we can guarantee that a boundary exists to separate clips that contain the target event from those that do not. Compared to other weakly supervised learning methods (Hospedales et al., 2011; Xu et al., 2013), the proposed approach has a much lower complexity, resulting in computational efficiency and robustness to the parameter initialisations; offers more utility than unsupervised approaches in that specific events can be detected; and greater practicality compared to supervised approaches due to the greatly reduced annotation demands. Furthermore, the requirement that the target events are rare matches many real world surveillance applications that focus on security.[1]

[1] This work extends that presented in Xu, Denman, Fookes, and Sridharan (2015), providing a much richer explanation of the proposed approach and a more thorough evaluation.

The use of KL divergence is motivated by the seminal work of Wang et al. (2009), who first proposed the use of probabilistic topic models for activity recognition. As part of this work, Wang et al. (2009) showed how semantic queries could be constructed by labelling the learned topics with a corresponding "real-world" definition. Distributions pertaining to each topic could then be compared to distributions for input clips using KL divergence to determine if an incoming clip matched an input semantic query. While this approach provides the motivation for our work, our work is significantly different in that we use KL divergence and weakly supervised learning to detect target events of interest automatically.

We note that while other techniques from the field of information theory have been proposed for crowd behaviour analysis (Gu, Cui, & Zhu, 2014; Zhao, Yuan, Su, & Chen, 2015) , these techniques have been concerned with detecting "scene wide" abnormal events (i.e. a rapid escape) using one-class unsupervised learning. The approach proposed in this paper is targeted at detecting specific known events which are embedded into a set of background activities using weakly supervised learning. However, we also demonstrate how the proposed approach can be used for abnormal event detection to further illustrate its utility.

The remainder of this paper is organised as follows: Section 2 outlines the proposed approach; Section 3 presents an evaluation of the proposed approach; and Section 4 concludes the paper.

2. Proposed approach

The proposed approach extracts bag of words features from short video clips (see Section 2.1), and models these using KL divergence (see Section 2.2) to detect specific events of interest. We also show how the proposed approach can be used to detect abnormal events, rather than specific events of interest (see Section 2.3).

The proposed approach uses the bag-of-words paradigm, originally from the field of natural language processing. Within this framework, a visual word is a quantised feature, and the entire set of visual words forms the vocabulary. A small video clip, or *document* can then be transformed into a multi-set bag of words, represented as a histogram. An event is modelled as a stationary stochastic process that is defined on the vocabulary (i.e. the underlying probability distribution of the words is constant). In our approach, there are two general types of event: the *event of interest* is the event that we are trying to detect; while a *background event* is any other event.

2.1. Feature representation

We consider two feature representations: the discrete optical flow approach of Wang et al. (2009), and a trajectory representation similar to that proposed by Xu et al. (2013). Within the trajectory based approach, we consider three approaches to build the trajectories: MPEG motion vectors as in Xu et al. (2013), particle video as used by Xu, Denman, Sridharan, and Fookes (2012b), and the KLT (Kanade–Lucas–Tomasi) tracker (Shi & Tomasi, 1994).

2.1.1. Discrete optical flow

The *discrete optical flow* descriptor (Hospedales et al., 2011; Wang et al., 2009) encodes the moving pixel's location and velocity into a discrete codebook, where the velocity is computed using optical flow. Typically, this feature descriptor is used alongside probabilistic topic models (Blei, Ng, & Jordan, 2003; Wang et al., 2009). This feature was used in the weakly supervised joint topic model of Hospedales et al. (2011), and descriptors were derived using a patch based approach which encoded the mean of the