Pattern Recognition Letters 49 (2014) 24–32

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

A nonparametric approach to region-of-interest detection in wide-angle views ☆

Guangchun Cheng *, Bill P. Buckles

Dept. of Computer Science and Engineering, University of North Texas, Denton, TX 76203, USA

ARTICLE INFO

Article history:
Received 26 March 2013
Available online 16 June 2014

Keywords:
Region of interest
Anomaly detection
Structure tensor
Unsupervised learning
Video analytics

ABSTRACT

We propose a tracking-free method to detect the regions of interest (ROI) in a wide-angle video stream. A region is defined as a statistical outlier among occurrences of motion patterns, and is detected in an unsupervised manner. Based on 3D structure tensors, the activity at any site is modeled by the probability distribution of distances between structure tensors. The distribution is estimated using a nonparametric kernel density estimator. The detection of regions is determined by observing a long period of low-probability motion occurrences. Experiments performed with real-world datasets indicate that the proposed algorithm can detect both spatial ROIs and spatio-temporal ROIs, and outperforms other nonparametric methods.

Published by Elsevier B.V.

1. Introduction

Efficient region of interest identification has been an active topic in computer vision fields such as visual attention in images and anomaly detection in video sequences. Currently the cameras record sufficient visual information for monitoring purposes. In fact, in most instances it is impractical for either human observers or automated systems to analyze each pixel in detail. Therefore selective operations are needed at different sites in a visual scene. Through region-of-interest (ROI) detection, non-interesting (e.g. normal) events can be excluded and further explicit event recognition methods can be applied to the remainder. This mimics the primates' visual system. However, it is not a trivial problem for man-made systems to understand the scenes and perform such selective processing.

The localization of ROIs becomes more urgent when given wide-angle camera views with background motion clutter. Most surveillance videos are produced in this manner to obtain a large and efficient coverage of a monitored area. The challenge we address is that the greatest portion, both spatially and temporally, of the video is not of interest. Traditional approaches such as tracking-based methods are not efficient, especially in clustered scenarios. Tracking-based approaches perform well in narrow-angle views or sparse scenarios with limited objects. In wide-angle views much more information must be processed which lowers the efficiency. More importantly, because it is not goal-driven, much computation is needed to establish which are the "normal" trajectories or other representations. Therefore, some researchers have begun applying methods that first localize the regions of interest, followed by operations such as tracking and anomaly recognition. This work is also motivated by this framework. The focus is on identifying potential ROIs in videos for further analysis.

Region of interest detection is basically a classification problem for which visual information is assigned labels of "interesting" and "non-interesting". For local feature representation, the description can be descriptive (such as common trajectory) or probabilistic (such as histogram). Correspondingly, the identification of local interest is based on the distance or probability of an observation compared with the canonical description. The relationships among information of different sites are also exploited in some probabilistic graphical models (e.g. conditional random fields). In order to model the information and detect the regions of interest, existing studies mainly use local information to model the activities [14,26,16]. Usually it is assumed that the statistical information follows a basic distribution (e.g. Gaussian) or a mixture of them. The training phase is designed to compute the parameters according to optimization criteria. It is not always straightforward to estimate the parameters and it is difficult to determine the form of the distribution or the number of the mixed models that should be applied to arbitrary videos. The innovations of the described method are given below.

- 3D structure tensors are used as the basis to extract tracking-free features to characterize the motion and spatial dimensions concurrently; bypassing object tracking avoids the computational expense and the errors it may induce.

☆ This paper has been recommended for acceptance by M. Tistarelli.

* Corresponding author. Tel.: +1 940 2979706; fax: +1 940 3698652.
E-mail address: guangchuncheng@my.unt.edu (G. Cheng).

http://dx.doi.org/10.1016/j.patrec.2014.05.019
0167-8655/Published by Elsevier B.V.

- A nonparametric approach models the distribution of tensor instances, treating observations with large deviation from the norm as statistical outliers to localize the regions of interest. This approach avoids the estimation of parameters as is required in parametric models.
- Characteristics of abnormal (or normal) spatial and motion patterns need not be explicitly specified; unsupervised training is applied to detect the norms and then the regions of interest.

Our first assumption is that the underlying processes that produce the motion change distribution are stationary and ergodic. That is, the mean of an observed sequence is an approximation of the mean of the population. While it is not difficult to exhibit non-stationary examples, we observe the motion changes at a specific site are most likely stationary and ergodic for extended periods. Switching to a new context, e.g., daytime activity vs nighttime activity, is a simple matter of reconstructing the motion pattern. For some types of videos such as movies, the interests are often defined by the *story* and *intent*, which fall outside the scope of this paper. From a bottom-up perspective of view, which seeks interests from features instead of goals, we also assume that interesting events are rare although the converse may not be valid. Our approach is to mark interesting events and allow for further video analytics to classify.

The rest of the paper is structured as follows. In Section 2, a brief review on anomaly detection in videos is given with the focus on non-tracking methods. Section 3 then explains in detail our approach to detect the regions of interest. Section 4 verifies our approach through experiments. The conclusion is in Section 5 with a discussion of limitations and future work.

2. Related work

There has been much research on anomaly detection in computer vision and other fields. ROIs in videos consist of two categories: spatiotemporal outliers and task-driven occurrences. Task-driven ROIs are determined by not only the low-level features but also the observer's tasks and context information [33]. Although task-driven approaches are appealing, the current focus is on the detection of spatiotemporal outliers. In this brief review, we focus on spatiotemporal outlier detection with the aim of pixel-level anomalous detection in videos, especially those with wide-angle views.

The representation of events basically includes object trajectories and functional descriptors. In a trajectory-based representation, objects are detected, tracked, and the trajectories are used to model events. Those trajectories deviating greatly from the more frequent ones are deemed outliers/anomalies [13,28]. In a functional descriptor-based representation, static background and background motion are modeled by functions using methods such as mixture of Gaussians [10] and kernel density estimation [21].

There exist two different paradigms for video anomaly analysis, namely (A) event recognition followed by anomaly determination and (B) (potential) anomaly detection followed by event analysis. In paradigm A, most anomaly detection systems are based on tracked trajectories associated with the object's speed and other properties [36,27]. Then the extracted features are used to develop the models in a supervised manner. Piciarelli et al. [25] used a single-class support vector machine (SVM) to cluster the trajectories for anomalous event detection. Morris and Trivedi [23] used a hidden Markov model to encode the spatiotemporal motion characteristics, and abnormal trajectories were detected. By defining and comparing the similarity between trajectories, [38] proposed a framework for anomaly detection in different scenes. A general weakness of paradigm A is that the anomaly detection accuracy depends on the results of tracking or event recognition. For wide-angle views such as surveillance videos, anomalies are rare, so tracking each object is not necessarily needed.

In the second paradigm, a potential ROI is first detected. Saligrama et al. [26] proposed a strategy that detects ROIs prior to higher-level processing such as object tracking, tagging and classification. It avoids unnecessary processing caused by assessing normal recurring activities by constructing a behavior image, a.k.a background activity map. Piciarelli and Foresti [24] compared explicit event recognition and anomaly detection, then combined both for surveillance anomaly detection. Gong and Xiang [8] recognized scene events without tracking with the aid of pixel change history (PCH). Zhong et al. [37] borrowed an idea from document-keyword clustering and used a co-occurrence matrix of spatial motion histograms to detect unusual activities in videos. Other research on anomaly detection include [35] who modeled background activity with moving blobs, [15] who developed HMMs with spatiotemporal motion patterns, and [14] who used MRFs to detect both local and global abnormalities, among others.

Some of the methods above are parametric while the others are nonparametric. Parametric methods such as MRF [14] and HMM [16] have been used by many. Although they are concise and more precise if the assumptions are correct, parametric approaches generally cannot be robustly applied to different scenes without modification. It involves model selection, estimation of the model order, and the model parameters. Gong and Xiang [8] provide an excellent example of estimating both the model order and the parameters. There are also many nonparametric approaches, most of which use local information and detect anomalies in a bottom-up manner. Based on the framework of image visual saliency detection [12,11] developed a model that computes the degree of low-level surprise at each location, which allows more sophisticated analysis within the most "surprising" subsets. Gaborski et al. [7] also extended the computational framework of [12] to identify inconsistent regions in video streams. As another typical nonparametric method, kernel density estimation [3] was used to detect network anomalies subject to a tolerance [2]. Laxhammar et al. [18] compared the Gaussian mixture model (GMM) with a kernel density estimator (KDE) using sea traffic data. They concluded that KDE more accurately characterizes features in the data but the detection accuracy of the two models is similar. Surveillance videos exhibit many variations in both the actions involved and the scene conditions, we base our work on nonparametric methods.

3. Pixel-level wide-view ROI detection

Video is commonly considered a sequence of frames $I^{(t)},\ t = 1, 2, \ldots, T$. Examiners usually can find objects or regions of interest from the sequence without any knowledge beyond the video. It is our hypothesis that outliers to the statistics within frames mark the ROIs. There are assumptions. First, normal activities outnumber anomalies. That is, statistical outliers correspond to regions of interest. Second, normal activities are sufficiently repetitive to form majority patterns which are the basis for statistical methods. Third, normal patterns have a finite lifespan. Changes in normal patterns, i.e. "*context switches*", are not covered by this work. These assumptions are common in cases such as traffic, crowds, and security zone surveillance.

The framework is illustrated in Fig. 1. We first compute a 3D structure tensor to capture the motion at the sampled site $\vec{x} = (x, y)$ for each frame $I^{(t)}$. Next the probability distribution of structure tensor is estimated in the online training phase using the structure tensor's eigenvalues. This is followed by the interest point detection as occurrences with low probability. ROIs are obtained using filtered interest points.