Neurocomputing 187 (2016) 66–74

Contents lists available at ScienceDirect

Neurocomputing

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

Event-based large scale surveillance video summarization

Xinhui Song [a], Li Sun [a], Jie Lei [a], Dapeng Tao [b], Guanhong Yuan [a], Mingli Song [a,*]

[a] College of Computer Science, Zhejiang University, Hangzhou 310027, China
[b] School of Information Science and Engineering, Yunnan University, Kunming 650091, Yunnan, PR China

ARTICLE INFO

Article history:
Received 7 July 2015
Received in revised form
24 July 2015
Accepted 26 July 2015
Available online 12 December 2015

Keywords:
Surveillance video
Large scale video summarization
Key area reconstruction

ABSTRACT

Recent advances in sensor manufacture and computer vision technologies have simulated the applications of intelligent transportation systems, while a key yet under-addressed issue in these systems is the semantic summarization of large scale surveillance video. The main difficulty of large scale surveillance video summarization arises from the contradiction between the high-degree spatiotemporal redundancies and the limited storage budget. In this paper, we propose a novel approach of large scale surveillance video summarization on the basis of event detection. In the proposed approach, we firstly obtain the trajectories of vehicles and pedestrians in a tracking-by-detection manner, and then detect the abnormal events using the trajectories. Finally, we design a disjoint max-coverage algorithm to generate a summarized sequence with maximum coverage of interested events and minimum number of frames. Compared with traditional key frame-based approaches, our approach enjoys the following favorable features. First, important information can be efficiently extracted from the redundant contents since the approach is event-centric and those interested events contain almost all the important information. Second, abnormal events are successfully detected by combining the Random Forest classifier and the trajectory features. Third, the abnormal events are designed to display, and hence further reduces the compression ratio. Due to the above features, the proposed approach is suitable for different scenarios, ranges from highway to crowded crossings. Experiments on 12 surveillance sequences validate the effectiveness and efficiency of the proposed approach.

1. Introduction

Recent advances in sensor manufacture and computer vision technologies have simulated the applications of intelligent transportation systems with millions of surveillance cameras capturing traffic data every day. Although bringing convenience to human life, the big surveillance data raise the challenge of time and space consuming when retrieving the video. Hence it is necessary to gain certain perspectives of the data without watching the whole video, and video summarization techniques are proposed to handle the problem.

Existing video summarization approaches use key frames to represent the videos as summarization, and most of them [1–5] rely on shot boundary detection. However, in the surveillance data, there are no explicit shot boundaries. Thus it is difficult to decide which frame is more important than others. As a consequence, the summarization results of key frame-based methods are not satisfied for surveillance videos. On the other hand, the key frames selected from the original videos may only has a small area with information, while other areas have no valuable information. Hence, the expressivity of key frame-based methods is limited.

In this paper, we propose a novel event-based approach for surveillance video summarization. In contrast with the shot-centric method, our approach is event-centric. First, the proposed approach does not rely on shot boundary detection. Second, since people put much more concerns in the abnormal events than the routines in a surveillance video, our approach works better in extracting the important information from redundant contents. Third, we propose the disjoint max-coverage algorithm and achieve higher summarization ratio than the key frame-based methods.

Specifically, the proposed approach consists of three steps. First, the trajectories of pedestrians and vehicles are obtained by combining the deformable part-based detector and multi-cue data matching in a tracking-by-detection framework. Second, abnormal events in the surveillance videos are detected based on the extracted trajectories. We propose trajectory features, which are discriminative for abnormal event detection when combining with Random Forest classifier. Third, a disjoint max-coverage algorithm is proposed to generate a summarized sequence with maximum coverage of interested events and minimum number of frames. By this means our approach achieves a very high summarization ratio.

* Corresponding author.

http://dx.doi.org/10.1016/j.neucom.2015.07.131
0925-2312/

In summary, the main contributions of this work are as follows:

- We propose the event-based video summarization approach, which is more suitable for surveillance video summarization than key frame-based approach.
- We combine the Random Forest and the trajectory features to detect abnormal events, and achieve good results.
- We design a disjoint max-coverage algorithm to generate a summarized sequence with maximum coverage of interested events and minimum number of frames.

The rest of the paper is organized as follows. Section 2 reviews the related work. Section 3 presents the abnormal events detection and summarization framework. Section 4 demonstrates the experimental results and Section 5 concludes the paper.

2. Related work

2.1. Abnormal event detection

Existing work on abnormal event detection can be roughly categorized into trajectory analysis and volume matching.

Trajectory analysis-based abnormal event detection approaches track the targets in the videos, stabilize the positions into trajectories, and then recognize whether they are abnormal events. Efros et al. [6] introduce a motion descriptor based on optical flow measurements in a spatiotemporal volume for the tracked human targets, and use an associated similarity measure in a nearest-neighbor framework to recognize the event. Jiang et al. [7] track all moving objects in the video and three different levels of spatio-temporal contexts are considered as features. In each level, frequency-based analysis is performed to automatically discover the rules of the abnormal events. Wu et al. [8] propose a method for detecting and localizing abnormal events in complicated crowd sequences using a Lagrangian particle dynamics approach, together with chaotic modeling. The method performs well for a range of crowded to sparse scenes. Hong et al. [9] proposed a novel method for the efficient object retrieval by applying the motion prediction in videos. And image processing [10–12] requires real-time response for processing abnormal events.

Volume matching-based abnormal event detection approaches apply a spatiotemporal volume localization scheme to search for the position of abnormal events. Ke et al. [13] propose to use the volumetric features for video event detection. Spatiotemporal shapes are correlated to video clips, and when combined with flow-based correlation technique, the approach can detect a wide range of actions in video. Boiman et al. [14] propose ensembles of patches to detect irregularities in videos. They pose the problem of determining the validity of visual data as a process of composing a new observed video segment using chunks of data extracted from previous visual examples. Regions in the observed segments which cannot be composed from the database are regarded as abnormal event. And the problem is formulated as an inference process in a probabilistic graphical model.

2.2. Video summarization

Video summarization techniques aim at summarizing the video with rich information into a small number of frames without losing the information.

Most existing video summarization techniques are key frame-based, i.e., several frames from the original videos are extracted to represent the whole video. E.g. Divakaran et al. [2] firstly divide the videos into equilong segments and regard each segment as a shot, and extract the first and the middle frames of the segment as the key frames. Because of its simplicity, this algorithm suffers from redundant contents in the video. Zhang et al. [3] use color histogram as the feature of each frame, and select a key frame if it is significantly differs from the previous chosen one. However, the algorithm cannot guarantee the representativeness of the chosen frames considering the redundancy. Morere et al. [15] propose to combine deep convolutional neural networks and restricted Boltzmann machines for key frame-based summarization. An original co-regularization scheme is used to discover meaningful subject-scene associations and the resulting multimodal representations are used to select highly relevant key frames. Lee et al. [16] present a video summarization approach, which produces a compact storyboard summary for egocentric camera. The resulting summary focuses on the important object and people that the camera captures.

Cernekov et al. [4] proposed to firstly detect shots in the video, and then extract key frames based on the mutual information and the joint entropy. Kelm et al. [5] detect gradual and abrupt cuts to segment the video into shots, and then extract key frames using visual attention features. Liu et al. [1] propose to jointly detect the shot boundaries and extract the key frames in a probabilistic framework, and use Gibbs sampling to infer the model. Luan et al. [17] select highlight candidate frames from each shot, and then reconstruct the candidate set by nonnegative linear reconstruction. The above methods rely on effective shot boundary detection algorithm, which are impractical in surveillance video.

Zhuang et al. [18] propose clustering-based key frame summarization. They first extract color and texture features to represent each frame, and then utilize K-means to obtain several clusters. Finally, if the frame number in a cluster is larger than a threshold, a frame is extracted as the key frame of this cluster. Similarly, Girgensohn et al. [19] also use clustering-based summarization. They require that each key frame should have a neighbor frame that belongs to the same cluster as the key frame, and the time interval between the two frames is less than 9 s. The similarity between frames is based on maximum matching [20] and optimal matching bipartite graph matching [21,22]. The clustering-based methods usually ignore the temporal information, which is very important in surveillance video summarization.

Cong et al. [23] provide a content summarization method as a sparsity consistency-based dictionary selection problem, in which a dictionary of key frames is selected such that the original video can be best reconstructed from the dictionary. Sze et al. [24] propose a global statistics-based key frame extraction scheme, where each pixel in the key frame is constructed by considering the probability of occurrence of those pixels at the corresponding pixel position among the frames. Khosla et al. [25] introduce prior information from web images to summarize videos containing a similar set of objects. The above reconstruction-based algorithms suffer from the same problem with clustering-based algorithms, hence the temporal information is lost.

3. Approach

Given surveillance videos, we first apply object tracking-by-detection to obtain the trajectories of pedestrians and vehicles, and then use Random Forest to recognize the abnormal events from the trajectories. Finally, a disjoint max-coverage algorithm is proposed to reconstruct the abnormal events. The flowchart of the proposed approach is illustrated in Fig. 1.