Neurocomputing 171 (2016) 347–361

Contents lists available at ScienceDirect

Neurocomputing

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

Unsupervised discovery of crowd activities by saliency-based clustering

Tingting Han, Hongxun Yao*, Xiaoshuai Sun, Sicheng Zhao, Yanhao Zhang

School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China

ARTICLE INFO

Article history:
Received 11 December 2014
Received in revised form
19 April 2015
Accepted 23 June 2015
Communicated by
Available online 29 July 2015

Keywords:
Spatio-temporal saliency
Unsupervised discovery
Crowd activity analysis
Spatio-temporal pooling
Dynamic time warping
Normalized cut

ABSTRACT

Along with the rapid development of digital information technology, video surveillance systems have been widely used in numerous public places, such as squares, shopping malls and banks, to monitor crowd in case of anomalous events. Meanwhile, great challenges have been posed to worldwide researchers because the analysis of the exponentially growing crowd activity data is an arduous task. In this paper, we develop a novel unsupervised crowd activity discovery algorithm aiming to automatically explore latent action patterns among crowd activities and partition them into meaningful clusters. Inspired by the computational model of human vision system, we present a spatio-temporal saliency-based representation to simulate visual attention mechanism and encode human-focused components in an activity stream. Combining with feature pooling, we can obtain a more compact and robust activity representation. Based on affinity matrix of activities, N-cut is performed to generate clusters with meaningful activity patterns. We carry out experiments on our HIT-BJUT dataset and the UMN dataset. The experimental results demonstrate that the proposed unsupervised discovery method is fast and capable of automatically mining meaningful activities from large-scale and unbalanced video data with mixed crowd activities.

1. Introduction

Unsupervised learning plays an important role in knowledge exploration and discovery. Since the increasing amount of multimedia data brings more and more laborious labeling work, it becomes a hot trend to apply unsupervised methods to handle relevant problems in computer vision research. For the last decades, unsupervised approaches have been extensively studied for object localization [1] and segmentation [2], action categorization [3,4], human activity analysis [5] and facial image analysis [6], etc.

Meanwhile, crowd activity analysis has attracted significant research interests in recent years for its potential applications in intelligent security monitoring of public places, such as railway stations, shopping malls and crowded sports arenas. Most of existing works on crowd activity analysis [7–9] are based on supervised learning on labeled data and detect abnormal activities by matching or classification strategies. However, there are some issues of these methods: (1) the supervised approaches are appropriate only when there are abundant labeled training samples, which is obviously burdensome and impractical; (2) even if adequate samples are given, it is still not flexible to build one single model for accurate inferences on large scale and diverse data stream.

To overcome these limitations, we attempt to automatically cluster a set of unlabeled crowd activity videos based on similarity affinities and discover semantically meaningful structures for unsupervised crowd activity analysis. It is of great significance to perform unsupervised analysis for crowd activities. First of all, it is an alternative approach to use unsupervised techniques to discover or mine visual patterns potentially existing in the huge amount of crowd activities. While traditional supervised methods may fail because it is unrealistic to enumerate all possible types of crowd activities which possibly contain interactions of various numbers of people or objects, occur in different scenes and involve many kinds of events. Secondly, providing a feasible way to label or annotate the crowd activity videos, it is conductive to liberate manual labor and can serve as a preprocessing for crowd activity analysis in supervised manners. Besides, we can better organize the data and realize searching in it. Last but not least, the proposed unsupervised analysis method is a general framework and thus can be also applied to analyze other types of activity or action video data.

However, it is still rather challenging to analyze crowd activities in a non-supervised manner. Unlike action analysis of a single person, the understanding of activities performed by multiple or a crowd of individuals has to overcome thorny problems such as diverse semantics, various expressions, complex interactions, occlusions and

* Corresponding author.
E-mail address: h.yao@hit.edu.cn (H. Yao).

http://dx.doi.org/10.1016/j.neucom.2015.06.048
0925-2312/© 2015 Elsevier B.V. All rights reserved.

low resolution. Therefore, in order to better describe activities, an effective and robust feature representation is required. Secondly, the scale of the realistic crowd activity data is very large and as a result, the proposed approach should be fast enough to analyze so many videos in an acceptable period of time. Thirdly, in the real world, normal activities usually account for the largest proportion of all the surveillance videos, while abnormal activities only make up a tiny proportion. As a consequence, the clustering or mining approach should be capable to overcome the intrinsic imbalance of the data.

Faces with the challenges, the contributions in this paper are threefold:

- We present a spatiotemporal saliency strategy to locate the human-focused dynamic changes within the stream of activities inspired by the computational model of human vision system, which can comprehend activities from the perspective of human. It is worth noting that, because it abstracts only important and distinctive information and eliminates mass of redundant information, our unsupervised discovery method has achieved high effectiveness and time efficiency.
- We introduce the graph-theory based clustering method, named Normalized-cut, and achieve high effectiveness for highly unbalanced data benefiting from its inherent advantages of solving the clustering of non-convex sphere of sample spaces and converging to global optimal solution.
- To support our unsupervised crowd activity analysis and to simulate the imbalance property of realistic data, we construct a new dataset named HIT-BJUT containing various types of human activities which is recorded in the campus of BJUT and will be released soon.

One preliminary version on unsupervised discovery of crowd activities was first introduced in our previous work [10]. In this paper we have improvements in three aspects: (1) we perform a more comprehensive survey of related works; (2) we enrich the theory of the proposed method; and (3) we conduct more comparative experiments and provide more analysis of the results. The remainder of this paper is organized as follows. Section 2 reviews the related methods and features for crowd activity analysis. The proposed unsupervised crowd activity analysis method based on spatio-temporal representation is described in Section 3. The test datasets and evaluation criteria are presented in Section 4, followed by experimental results and analysis in Section 5. Finally, conclusions are drawn in Section 6.

2. Related work

Significant progress has been made for action recognition over the last decade [11,12]. By far, the most popular representation for RGB video is Bag-of-Visual-Words (BoVW) model based on low-level features, such as local interest points [13] and dense trajectories [14]. More recently, increasing efforts have been focused on exploring mid-level features. Discriminative action parts such as motionlets [15], actons [16] and attributes [17] have been mined to describe actions. With the development of the techniques for capturing depth data, researchers have also been exploring action representation of 3D actions [18,19].

While action recognition focuses on motion patterns of single person or pair-wise persons, crowd activity analysis takes the interaction of the crowd into account and treat it as a whole to represent, which is more complicated and challenging due to the problems of diverse semantics, various expressions, complex interactions and occlusions, etc. Recent works on crowd activities mainly focus on the topic of anomaly detection, which aims to discover the abnormal activities from a set of normal activities. Conventional methods tend to address the challenging task in a supervised manner, and can be coarsely divided into two categories: model-based methods and particle advection-based methods.

In the framework of model-based method, it is a convention to train a model for normal crowd activities, and those activities which cannot be covered by the trained model are identified as the anomalies. In literature [20], a dynamic texture model jointly modeling appearance and dynamic information in a crowd scene was employed to detect both temporal and spatial anomalies. Kratz et al. [21] analyzed the underlying structure formed by the spatial and temporal variations in the motion to exploit the steady state motion of crowd activities and modeled the motion patterns with a Hidden Markov Model (HMM). In this way, abnormal activities could be detected as the motion patterns with low likelihood. In [22], Kim et al. tackled the abnormal activity detection by spatial–temporal Markov Random Field (MRF). Amer et al. [23] presented a new deep model, called the Hierarchical Random Field (HiRF), for representing and recognizing collective activities in videos. Antic et al. [24] proposed a technique based on video parsing to achieve abnormality detection. Cong et al. [25] proposed to detect the presence of anomalies in crowded scenes by a sparse reconstruction cost. While more popular methods in recent years are based on particle advection schemes. In these methods, a grid of particles are considered in each frame which are then advected using the underlying motion data [7,9,26,27]. In [7], Social Force Model (SFM) was employed to detect abnormalities by estimating the interaction force, which in turn, was used to describe crowd behavior. Although these methods have obtained the state-of-the-art performances, the difficulties of labeling huge amount of training samples and modeling a single model for activities with complex semantics and diverse expressions still remain to be overcome.

In addition to the learning methods, another key point is to extract effective features from the spatio-temporal video data. There are usually two types of widely used features: (1) high-level features abstracted from the detected objects by tracking and recognizing process; (2) low-level features directly derived from the image pixels. Most commonly used object-based features are the position and trajectory of the object's centroid, which have been proved to be effective in detection of various types of anomalies, such as running and falling [28] and traffic anomaly [29]. In addition, features such as limb angles [30,31] can be used to identify the poses or actions performed by the objects when the resolution of the videos is high enough. However, to abstract these features, the separation and tracking of multiple targets have to be solved as a preprocessing stage. Unfortunately, these two problems themselves remain challenging because of the severe occlusion and other distractions existing in the videos. On the other hand, low-level features become more and more popular when dealing with abnormal activity detection benefiting from their robustness against the aspects, such as occlusion, which would affect the tracking accuracy negatively. Whats more, the abstraction of low-level features does not rely on the segmentation or tracking of objects, so it makes them effective when there are large numbers of targets in view. As the most basic and successful feature to describe motion information in videos, optical flow is widely used in the task of anomaly detection [7,32,33]. Features such as color, texture, gradient and shape have also been explored for this purpose [20,34–36].

Different from these methods, we develop an unsupervised activity analysis method to explore the underlying patterns existing in the data without involving manually labeling and complex modeling. Moreover, we propose a saliency-based feature map which can abstract information inconsistently capturing more important components that humans focused on and ignoring those of little significance which would cause unnecessary