Computer Vision and Image Understanding 144 (2016) 177–187

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

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

Spatio-temporal texture modelling for real-time crowd anomaly detection

Jing Wang, Zhijie Xu*

Visualisation, Interaction and Vision Research Group, University of Huddersfield, United Kingdom

ARTICLE INFO

Article history:
Received 20 December 2014
Accepted 28 August 2015

Keywords:
Crowd anomaly
Spatio-temporal volume
Spatio-temporal texture

ABSTRACT

With the rapidly increasing demands from surveillance and security industries, crowd behaviour analysis has become one of the hotly pursued video event detection frontiers within the computer vision arena in recent years. This research has investigated innovative crowd behaviour detection approaches based on statistical crowd features extracted from video footages. In this paper, a new crowd video anomaly detection algorithm has been developed based on analysing the extracted spatio-temporal textures. The algorithm has been designed for real-time applications by deploying low-level statistical features and alleviating complicated machine learning and recognition processes. In the experiments, the system has been proven a valid solution for detecting anomaly behaviours without strong assumptions on the nature of crowds, for example, subjects and density. The developed prototype shows improved adaptability and efficiency against chosen benchmark systems.

1. Introduction

The increasing demands of intelligent surveillance applications have triggered many innovative developments in automated video event detection areas. These techniques, such as crowd anomaly detection, can be used in monitoring and tracking emergency situations occurred in crowded scenes, including busy motorways, high streets, sporting events, and open air concerts.

Modelling crowd scenes from video recordings present tough challenges due to severe occlusion problems among crowd subjects. Traditional top-down approaches, which focused on accurately tracking and identifying the so-called "abnormal" behaviours of crowd entities, have been proven inefficient and inaccurate for crowd-based anomaly detection [1]. Crowd scenes often contain uncertainties such as changes of subject density, average subject size, shape, and boundaries of these entities that can bring ambiguities to the definition and interpretation of the meaning and natures of crowd anomalies. For tackling those problems, it has been widely acknowledged that the modelling of all (or most of) the crowd behaviours through analysing their constituent elementary characteristics in a chosen feature space is unavoidable (the so-called bottom-up process).

While the visual appearance of an individual subject's behaviour in a crowd scene may vary, their intra-/inter- group dynamics within certain quantifiable feature space are often statistically identical, for example, among a flock of birds, a school of fishes, and even vehicles travelling on a motorway. The swarm behaviours are constructed of similar "visual traces" for a studied scene. In the case of crowd anomaly analysis, such as to detect people's sudden gathering or dispersion, complex interactions of crowd subjects may change the visual appearance of the local/global image observations, which has motivated this research to model the crowd scene "changes" through analysing their overall grouping dynamics and facilitating real-time (or near real-time) anomaly warning applications.

It is worth noting that "normal" and "abnormal" events are intrinsically ambiguous on semantic level. Certain crowd behaviours are normal in one scenario but may become hazardous in others. For example, crowds running in a marathon are "normal", but groups of people suddenly start running in a shopping mall may indicate an accident or incident. Based on the nature of main stream surveillance applications, the occurrence of anomaly events usually counts a very small percentage of the entire surveillance cycle and often demands immediate verification and response. In this research, it is considered reasonable to define normal crowd behaviours as the dominate pattern. Instead of composing complex event models for semantic interpretation, a normality crowd model in this research can be learnt and self-updated by abstracting the visual features along surveillance time-lines.

The core research problem highlighted in this research is to abstract effective visual features which can accurately describe normal crowd activities within the spatial and temporal domain, and to devise a robust decision making algorithm for real-world settings. Due

* Corresponding author. Fax: +44 1484422252.
E-mail address: z.xu@hud.ac.uk (Z. Xu).

http://dx.doi.org/10.1016/j.cviu.2015.08.010
1077-3142/

to the application-oriented nature of this research, the implementation strategy should also be easily adopted by modern closed-circuit television (CCTV) systems with minimum lag.

In this paper, a crowd anomaly detection algorithm has been introduced based on image textures formulated by spatio-temporal information. The so-called spatio-temporal texture (STT) has been proven as an effective dynamic representation media and mechanism for single human action detection. This research explores its characteristics in maintaining the statistical consistency across crowd events domain and its sensitivity to group anomalies (or sudden changes). A "redundancy" feature space has been formulated based on the STT structure through wavelet transform, which enables a flexible multi-criteria binary decision making mechanism to be constructed.

This paper is structured as follows: a literature review of the related crowd modelling techniques have been introduced in Section 2. Section 3 focuses on the design of STT-based crowd modelling techniques. An innovative crowd anomaly detection algorithm is presented in Section 4. System evaluations over benchmarking approaches have been highlighted in Section 5. Section 6 concludes the research with a discussion on the pros-and-cons of the devised method and envisaged future improvements.

2. Literature review

Computer vision-based crowd behaviour analysis has attracted immense attention from the research and application domains since the 1990s [2].Various algorithms and techniques for assisting the precision and speed of the processes have been studied in the last two decades. In general, there are three representative approaches: individual feature-based, flow field-oriented, and spatio-temporal feature-driven.

In the first category, crowd behaviour is often treated as an assembly of individual activities aggregated from each crowd entity. For example, a crowd movement along a busy street can be recognised as a group of people walking in the same direction. These methods are focused on describing crowd attributes by locating, isolating and analysing each crowd member. Benefited from recent development on machine learning theories and practices, those methods, such as tracking pedestrians through face detection [3] and estimating crowd size through head contour counting [4], have been proven as powerful individual-oriented tracking strategies [2,5].

Methods in the second category define the crowd scene as a dynamic flow field, which is the most popular approach to date, for analysing crowd features. Early studies, such as the "Minkowski fractal dimension" model [6], and the flow-based "crowd motion" model [7], had focused on the extraction of crowd attributes from the vector fields to describe crowd density, moving directions, and boundaries. In recent years, more attention has shifted towards the application-oriented techniques to improve crowd pattern interpretations [8–10]. In 2007, Ali [11] first introduced a crowd scene model based on "finite time Lyapunov exponent field" - an extension of the flow-filed model - for segmenting extremely dense crowd scenes recorded in a video. The segmentation outputs are then used in the so-called "floor field model" calculation for tracking specific individuals from dense human crowds [12]. This model has also been applied in group tracking that containing multiple or intersected crowd entities [13]. Rodriguez's off-line dominating crowd moving direction learning algorithm [14] has also been proven as an effective flow-based tracking approach. Related research works, such as Crowd Kanade-Lucas-Tomasi (KLT) corners [15], multi-label optimisation [16], and Lagrangian particle trajectories ("work-flow" model) [17], have defined crowd anomaly features from abstracted flow-field data. Those methods demonstrated their potentials in tracking the dynamic crowd under extremely crowded and partial occluded conditions but are bound to pre-defined crowd patterns and specific applications

Different from the first category, the flow field-based techniques are mainly based on the so-called "global" crowd motions. The impact of the "local" and individual crowd member is often ignored. However, it is widely accepted that acts from localised entity or entity group can bring significant changes to a crowd's consistency and its dominant motion pattern. Therefore, a more generic crowd model should be established for describing the interactions of both the local and group entity features.

The third significant approach defines the crowd scene videos as Spatio-temporal Volume (STV), which combines global crowd dynamics into a three-dimensional model. For example, in 2009, Benezeth [18] introduced a motion labelling method based on the co-occurrence of features defined in STV. The model has been implemented as a function of the Markov Random Field (MRF) for crowd anomaly detection. Kratz [19] introduced a STV-based motion pattern modelling method for highlighting the spatial-temporal statistical characteristics of extremely dense crowd scenes. In 2012, Bertini [20] developed a STV-based anomaly location detection technique through using localised cuboids in an unsupervised learning framework. The method has been proven as a valid approach for non-parametric modelling of spatio-temporal features.

Recently, another important "post-" processing model employing information fusion techniques has been introduced. Pilot researches, such as Mehran's "social force model" [21] and its optimised versions such as "interaction forces" [22,23] have been developed to serve as the preliminary assumptions for crowd behaviour patterns. The models require pre-defined conditions to be satisfied before operating, for example, the majority of a crowd should move towards the same target area, which restricts their flexibility in real-world applications.

It is also worth noting that the unsupervised (and some self-supervised) machine learning algorithms are becoming popular for detecting crowd anomalies. For example, Feng introduced an online self-organizing map (SOM) [24] to model crowd scenes, which keep updating its "normal" patterns by using new feeds. Jiang [25] used an unsupervised clustering algorithm to compare crowd members and their neighbours for identifying significant variations – sign of anomalies.

In this research, an innovative anomaly crowd detection strategy based on the statistical crowd features extracted from STTs has been devised. The texture model contain strong statistical characteristics for describing repetitiveness and randomness of recorded scenes, which inherently combines local and global crowd entity features. It is based on the assumption that a crowd scene and its related dynamic patterns can be encapsulated in the texture model conforming human perceptual intuition. Whilst other approaches have attempted to combine crowd scene texture features with spatio-temporal information [26,27], the proposed method justifies its core value through enabling real-time anomaly detection without the time-consuming machine learning process and rigid assumptions on crown behaviour patterns.

3. Spatio-temporal texture modelling

Spatio-temporal Texture (STT) model is a statistical model developed in this research, which is sensitive to the sudden changes of crowd motions.

Fig. 1 illustrates the main steps for defining STT from raw video data. In this pipeline, STT is composed by using spatio-temporal volume and its slices. The slices contain crowd scene information and can be abstracted using wavelet transforms. After transforming each slice from spatio-temporal domain into wavelet space, the crowd