Applied Soft Computing 25 (2014) 485–495

Contents lists available at ScienceDirect

Applied Soft Computing

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

Multiresolution semantic activity characterisation and abnormality discovery in videos

Luis Patino*, James Ferryman

School of Systems Engineering, University of Reading, RG6 6AY, United Kingdom

ARTICLE INFO

Article history:
Received 28 June 2013
Received in revised form 24 June 2014
Accepted 18 August 2014
Available online 24 August 2014

Keywords:
Video understanding
Semantic multimedia extraction
Activity reporting
Human behaviour
Scene topology discovery
Abnormality detection

ABSTRACT

This paper addresses the issue of activity understanding from video and its semantics-rich description. A novel approach is presented where activities are characterised and analysed at different resolutions. Semantic information is delivered according to the resolution at which the activity is observed. Furthermore, the multiresolution activity characterisation is exploited to detect abnormal activity. To achieve these system capabilities, the focus is given on context modelling by employing a soft computing-based algorithm which automatically enables the determination of the main activity zones of the observed scene by taking as input the trajectories of detected mobiles. Such areas are learnt at different resolutions (or granularities). In a second stage, learned zones are employed to extract people activities by relating mobile trajectories to the learned zones. In this way, the activity of a person can be summarised as the series of zones that the person has visited. Employing the inherent soft relation properties, the reported activities can be labelled with meaningful semantics. Depending on the granularity at which activity zones and mobile trajectories are considered, the semantic meaning of the activity shifts from broad interpretation to detailed description. Activity information at different resolutions is also employed to perform abnormal activity detection.

1. Introduction

Advances in sensor devices, communication and storage capacities make it increasingly easier to collect large amounts of multimedia material. Recorded video and/or audio is then made available in standard databases. However, the value of this recorded data can only be exploited by technologies that can effectively transform this raw data into structured knowledge with rich semantics that can be naturally understood by an end-user. Specifically for long-term video monitoring applications, there is an increasing need for systems with automatic human activity characterisation and recognition and delivery of knowledge with semantics-rich content description.

However, in the field of video data-mining (including semantic video indexing and retrieval [1,2], mining frequently occurring objects and actions from videos [3–5], multimedia data mining for traffic video sequences [6,7] and mining of multimedia data for behaviour and activity analysis [8]), the gap between low-level image features and high-level semantic information still persists; and attempting to deliver semantic-rich description of the activities contained in the video remains a highly challenging task, not only because of the difficulty on automatically assigning semantic terms, but also because the semantic description of an activity could vary depending on the spatial context and the resolution level at which it is most appropriate to interpret the activity. A person waiting on a bus stop, for example, would be considered as a normal activity but if the person stands in another area it may be considered suspicious. In a broader view, if the person stands at different uncommon areas, it could be considered loitering.

This paper focuses on characterising and analysing the activity contained in the video in a multi-resolution way. The novelty consists in designing an unsupervised system for the extraction of semantic information on human activities from video where such semantic interpretation may vary following different resolution levels at which one observes the activity. Furthermore, abnormality detection is also performed based on the analysis of the characterised activity at different resolutions.

The activity characterisation is achieved through trajectory analysis. By employing clustering techniques, activity zones (context zones) can be learnt characterising the scene dynamics. People activities are extracted by relating mobile trajectories to the learned zones. The activity of a person can then be summarised as the series of zones that the person has visited. Zone-based

* Corresponding author. Tel.: +44 0118 378 5225; fax: +44 0118 975 1994.
E-mail addresses: j.l.patinovilchis@reading.ac.uk (L. Patino), j.m.ferryman@reading.ac.uk (J. Ferryman).

http://dx.doi.org/10.1016/j.asoc.2014.08.039

activity characterisation has already been proposed in the past [9] and shown to work successfully. However, delivering discovered activities with semantics and automatically detecting abnormal events has not been fully addressed in these approaches.

This work proposes to automatically attribute semantic labels to activity zones allowing to express activities themselves with semantics. Moreover, it is proposed to fully exploit the multiresolution information inherent to this type of activity characterisation for the detection of abnormalities. Two kind of abnormalities are considered: People staying too long on the observed scene and unusual taken paths.

Instead of learning specific models of normal/abnormal activity, statistical properties of zone occupancy and transition between zones are employed to infer statistical thresholds of abnormality. Detected mobiles can then be inferred as having a normal or abnormal activity. This work presents results from a number of different and representative surveillance domains and shows that the approach is viable.

The remaining of the paper is structured as follows. The next section gives a short overview of the related work. The employed activity characterisation approach is presented in Section 3. It is explained how to perform abnormality detection in Section 4, then the semantic labelling methodology is presented in Section 5. Section 6 gives the main results and evaluation. Finally, Section 7 draws the main conclusions and describes possible future work.

2. Related work

Activity from video usually corresponds to a predefined event recognition task. In this case the event model is manually defined and then encoded according to the chosen technology, for example, HMMs [10]; Bayesian classifiers [11]; constrained methods [12]; Soft computing [13]. The research trend is evidently to attempt to learn activity models either with unsupervised or with supervised learning. In the first case, techniques such as attribute multiset grammars/Bayesian Networks [14]; graph mining [15] have been proposed but in these cases they are limited to the interaction of two agents and few simple relationships between them. In the second case for supervised activity learning other different approaches have been employed such as Inductive Logic Programming (ILP) [16] or Context Free Grammars [17]. Such approaches have the down side that annotated data is often hardly available.

When the task comes down to abnormality detection, activity models are then also manually established to recognise the targeted abnormal event [18] or in most cases, anomaly detection equals an outlier detection process where 'abnormal data' does not fit normal activity patterns manually set or learned and often set within a Bayesian framework [19–21] but still either priors or thresholds indicating 'abnormally' low likelihood values are manually set. Sillito and Fisher [22] propose to allow operator intervention to incrementally learn a class of normal behaviour with a GMM classifier and cubic-spline points from human trajectories. Human intervention drops as the classifier correctly builds the model. Li et al. [23] analyse as well abnormal behaviour employing cubic-spline points from human trajectories but cast the problem in a complete supervised approach building a dictionary set of normal routes. Recently, a review on anomaly detection in surveillance has appeared where models are not only divided depending on the learning type employed (manually established, supervised/unsupervised learning) but also depending on their targets (traffic, individuals, objects, crowds) [24].

However, even when these approaches have been able to automatically build activity models in an unsupervised/supervised way for the application environment at hand, activity characterisation is still lacking semantic labelling so that an end-user can more naturally understand the extracted activities. Some research has been done in this aspect. Khan et al. [25] have predefined sets of actions, emotions, locations and objects; when these are recognised a template is employed to build grammatically correct sentences. Similarly Liu et al. [26], describe how to recognise predefined events in the subway, expressed in the form of an operation triplet given as subject–verb–object. Defining the object/subject corresponds to the task of detecting and classifying mobiles in the scene into human, baggage, train, train door. The semantic event recognition comes down to the verification of predefined rules between the object and the subject. Arens et al. [27] have built an activity recognition system where the activity is modelled with Situational Graphs and the recognised activity is outputted with some grammatically correct expressions. This is quite similar to the idea of having manually predefined activity event models and semantic tags which are activated when the event is recognised. In the same line is the work by Poppa et al. [28], designed to recognise predefined events in a shopping centre based on the combination of trajectory classification, action recognition and zone information (i.e. time spent in a given zone). The trajectory classification has three categories ('Disoriented', 'looking around' and 'goal oriented'). The number of behaviours that can be recognised is limited to 6 manual models ('looking for support', 'disoriented buying', 'disoriented', 'goal oriented buying', 'looking around' and 'buying').

The contribution of this work to the state of the art consists thus in building on a soft computing-based multiresolution activity characterisation to deliver appropriate semantic and extracting statistics from the activities at the different resolutions to infer statistical thresholds of abnormality. Detected mobiles can then be inferred as having a normal or abnormal activity.

3. Activity analysis

Activity patterns are based on the analysis of trajectories from detected mobile objects. Low-level tracking information is thus transformed into high-level semantic descriptions conveying useful and novel information.

The proposed system works offline once mobile objects have been detected and tracked. The hypothesis is that an online module has already performed the tracking process; hence low-level tracking is not considered in this work. The complete processing chain of the proposed approach is shown in Fig. 1.

The proposed system would then start by the analysis of detected mobile trajectories and extracting trajectory points of interest indicating mobile change of speed or direction.

As previously mentioned, complex activity recognition is achieved by learning the activity zones where mobiles evolve in the scene. This is the second step in the proposed system. The chosen zone-based approach is that given in [9] because of the possibility to characterise the activity and extract information at different granularities or resolutions. The zone learning procedure is thus only summarised in this paper. As observed from Fig. 1, the input to the zone learning procedure are the extracted trajectory points of interest. In a third step the mobile activity is characterised as a series of visited activity zones (activity extraction module). Such characterisation allows delivering behaviour events indicating the mobile activity. In a fifth step a check is made for abnormal activity patterns following the different established zone resolutions. Finally, all activity events are automatically labelled with semantics to give a more natural understanding of the ongoing activities to the end-user.