Pattern Recognition 41 (2008) 3481 – 3492

Contents lists available at ScienceDirect

Pattern Recognition

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

Recognising online spatial activities using a bioinformatics inspired sequence alignment approach

Daniel E. Riedel, Svetha Venkatesh, Wanquan Liu*

Department of Computing, Curtin University of Technology, GPO Box U1987 Perth, Western Australia 6845, Australia

A R T I C L E I N F O

Article history:
Received 25 May 2007
Received in revised form 16 April 2008
Accepted 17 April 2008

Keywords:
Activity recognition
Bioinformatics
Sequence alignment
Dynamic time warping

A B S T R A C T

In this paper we address the problem of recognising embedded activities within continuous spatial sequences obtained from an online video tracking system. Traditionally, continuous data streams such as video tracking data are buffered with a sliding window applied to the buffered data stream for activity detection. We introduce an algorithm based on Smith–Waterman (SW) local alignment from the field of bioinformatics that can locate and accurately quantify embedded activities within a windowed sequence. The modified SW approach utilises dynamic programming with two dimensional spatial data to quantify sequence similarity and is capable of recognising sequences containing gaps and significant amounts of noise. A more efficient SW formulation for online recognition, called Online SW (OSW), is also developed. Through experimentation we show that the OSW algorithm can accurately and robustly recognise manually segmented activity sequences as well as embedded sequences from an online tracking system. To benchmark the classification performance of OSW we compare the approach to dynamic time warping (DTW) and the discrete hidden Markov model (HMM). Results demonstrate that OSW produces higher precision and recall than both DTW and the HMM in an online recognition context. With accurately segmented sequences the SW approach produces results comparable to DTW and superior to the HMM. Finally, we confirm the robust property of the SW approach by evaluating it with sequences containing artificially incorporated noise.

1. Introduction

Video surveillance and automatic recognition of human activities is becoming increasingly important in modern society. Such recognition systems can be used to detect suspicious activities in complex environments such as airports and railway stations, to control interfaces in human computer interaction (HCI) applications or to provide a means for monitoring elderly individuals in a caring capacity. In this paper we recognise human activities, represented by spatial sequences, in a simulated smart home environment and further restrict the problem domain to the elderly. In doing so, we assume that the elderly carry out activities in a habitual manner; an assumption consistent with the findings of Monk et al. [1] and Suzuki et al. [2]. Providing smart houses for the elderly is important in maintaining the quality of life and independence of the aging population and reducing the on-going costs of care associated with that maintenance.

Activity recognition is the task of identifying an action or series of actions taken in pursuit of an objective. This view of activity recognition differs from existing works where researchers view the detection of walking, running or similar primitive actions as activity recognition. In a smart home setting, objectives could include but not be limited to having breakfast, reading a newspaper, watching television, cooking, having a shower or going to sleep. The series of actions that one would need to recognise in order to determine whether an objective has been accomplished may involve walking to a table, opening a cupboard, sitting on a chair or even lying on a bed. We focus our attention on the spatial component of activities as they can be obtained *non-invasively* from video tracking systems and for the majority of activities spatial signatures are unique.

Much work has been done in the area of human activity recognition over the last decade. Existing methodologies can be classified according to the manner by which activities are modelled, producing two distinct methodologies: state-space models and template matching techniques [3]. State-space models attempt to capture the variation in spatial sequences. These approaches include neural networks [4–6], hidden Markov models (HMMs) [7] and extensions to the HMM [8–12]. The HMM and its' variants have been used

* Corresponding author. Tel.: +61 8 92662746.
E-mail addresses: riedelde@cs.curtin.edu.au (D.E. Riedel), svetha@cs.curtin.edu.au (S. Venkatesh), W.Liu@curtin.edu.au (W. Liu).

0031-3203/$30.00 Crown Copyright © 2008 Published by Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2008.04.019

successfully in dealing with uncertainty but suffer from high training complexity, in particular the multi-layer and hierarchical models. Template matching approaches such as [13–17] compare extracted features to pre-stored patterns or templates, but have issues with high runtime complexity, noise intolerance, spatial activity variation, and/or viewpoint specificity.

The activity sequences used by these activity recognition approaches are typically captured using online video tracking systems. These systems produce continuous and uniform streams of tracking data that contain known activities and non-activity subsequences, corresponding to movement between activities and deviations from known activity paths. In order to isolate individual spatial activity sequences for quantification with models or templates one can either use a sliding window with width w on the buffered stream (Fig. 1(a)–(c)) or segmentation of the data stream. Sliding window approaches compare the fixed length window sequence to pre-stored templates for similarity quantification prior to classification, as shown in Fig. 1(d). Segmentation techniques recognise and extract embedded activities through location of activity boundaries, prior to similarity quantification and classification. Segmentation of continuous data is not a new problem and has been addressed previously in domains such as speech and gait recognition. Unfortunately, in the spatial activity domain segmentation is more difficult as activity boundaries are not so obvious.

Few methods have been proposed to specifically deal with online activity segmentation. In the methods of Bobick and Ivanov [18] and Ivanov and Bobick [19] a sliding window is applied to an observed sequence to allow for inferencing with low level HMMs. The observed sequence is then labelled accordingly. Like other sliding window approaches, the performance of this technique is sensitive to the specified window size. Another segmentation approach is given by Peursum et al. [20], where observed sequences are segmented and classified using HMMs trained with manually labelled activity sequences. During classification, the probability of a sequence having a particular label is determined and through calculation of the probability at each time instance, the boundaries of the activities can also be found.

To apply any of the above activity segmentation methodologies in isolation, for segmentation of a continuous data stream, is problematic. This is because the segmentation components of the approaches are intertwined with the recognition capabilities. Therefore, one must still adopt a sliding window approach to identify embedded activities, particularly if one wishes to use un-related sequence matching techniques for similarity quantification. Given this constraint, two issues relating to sliding windows need to be addressed. The first of these is the window size w. If one assumes that activities are conducted over a similar duration, as are habitual activities with the elderly, and in ideal tracking conditions (such as in controlled indoor environments) then it is appropriate to use a window size corresponding to the length of the longest activity. Realistically, the window size must be set to some value larger then the longest activity length, taking into account a feasible increase in possible activity duration. With an appropriately sized sliding window (normally set to the length of the longest activity or the average length ± 2.5 standard deviations), the second issue relates to locating an activity within that window sequence as shown in Fig. 1(d). Quantifying window sequences poses a problem for classification as the corresponding sequences can contain additional subsequence elements, which are not part of an embedded activity. These superfluous elements can in turn reduce the probability of an activity occurring in relation to a learnt model or increase the aligned distance between a class template. Even if duration is constant across activities, variation in captured sequence length occurs, due to tracking systems failing to consistently track objects. Some possible reasons for the failure result from occlusions, lighting variation, deficiencies in background subtraction techniques and geometric modelling limitations.

Techniques like the HMM [21] take the whole window sequence into account when calculating the probability of an observed sequence belonging to a given model. As a result, the superfluous elements decrease the resulting sequence probability, particularly if they have estimated symbol probabilities close to zero. Dynamic time warping (DTW) [22] and similar global sequence alignment approaches such as edit distance with real penalty (ERP) [23] and edit distance on real sequence (EDR) [24] are also susceptible to superfluous sequence elements. This occurs as the techniques attempt to minimise the distance across the entirety of both the known and observed sequences, taking into account the additional distance from the superfluous elements. Similarity algorithms based on the longest common subsequence (LCSS) [25,26] address this global limitation by ignoring superfluous elements in the observed sequences. Unfortunately, the techniques also allow significant deviations in a pattern, which can lead to incorrect classification. For instance, if one measures the LCSS between a known 1D sequence $\mathbf{a} = [13]$ and two observed sequences $\mathbf{b} = [123]$ and $\mathbf{c} = [1224443]$, both $\mathbf{b}$ and $\mathbf{c}$ have the same LCSS length (of three), indicating that both are equally similar to $\mathbf{a}$. However, by visual inspection, one would say that $\mathbf{b}$ is more similar to $\mathbf{a}$.

In order to recognise known spatial activity patterns and to address the above-mentioned deficiencies, we apply the Smith–Waterman (SW) local alignment approach from bioinformatics and modify the algorithm for two dimensional online spatial activity recognition. In previous work [27], we proposed the use of SW for spatial activity recognition and successfully evaluated the approach using accurately and inaccurately segmented activity sequences. In this paper, we provide OSW (online SW), a more efficient SW formulation for online recognition. Unlike DTW and HMMs, OSW does not require accurate sequence segmentation to correctly recognise embedded sequences, such as those found in sliding windows of online recognition systems. To prove the superiority of the OSW approach over DTW and the discrete HMM, we evaluate it in an online context with a sliding window and using a 12 activity data set. We also demonstrate the effectiveness of SW with accurately segmented activity sequences. The robustness claim of SW over DTW and the HMM is further validated by evaluating the approach with accurately segmented spatial sequences containing noise.

The layout of the paper is as follows. In Sections 2–4 we introduce sequence alignment, discuss the SW algorithm and the modifications for spatial sequence recognition, and then the proposed OSW algorithm. In order to perform a benchmark comparison of SW, we also provide a discussion of DTW (Section 5) and the discrete HMM (Section 6) and factors involved with their application in spatial activity recognition. Following from this, we present our data collection and experimental methodology in Section 7. Results from evaluation with a 12 activity data set are shown in Section 8 and a conclusion is presented in Section 9.

2. Sequence alignment

Sequence alignment methods are concerned with finding the best matching alignments of two query sequences according to specified optimisation criteria; typically maximising similarity or minimising distance. In order to derive the optimal alignments, each symbol is compared sequentially with the symbols of the other sequence. During this stage the local similarity or distance is calculated between the opposing symbols and using techniques such as dynamic programming (DP), optimal subalignments and finally an optimal alignment are produced. With the maximising similarity criteria a positive score is associated with matching symbols, while negative scores are given to non-matching symbols and insertions/deletions