Pattern Recognition ■ (■■■■) ■■■–■■■

Contents lists available at ScienceDirect

Pattern Recognition

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

Recognizing human motions through mixture modeling of inertial data

Matthew Field [a,*], David Stirling [a], Zengxi Pan [b], Montserrat Ros [a], Fazel Naghdy [a]

[a] School of Electrical, Computer and Telecommunications Engineering, University of Wollongong, Northfields Avenue, 2522 NSW, Australia.
[b] School of Materials, Mechanical and Mechatronic Engineering, University of Wollongong, Northfields Avenue, 2522 NSW, Australia.

ARTICLE INFO

Article history:
Received 21 December 2012
Received in revised form
3 January 2015
Accepted 5 March 2015

Keywords:
Human motion
Classification
Recognition
Segmentation
Inertial sensors
Gaussian mixture model
Minimum message length
Dynamic time warping

ABSTRACT

Systems that recognize patterns in human motion are central to improvements in automation and human computer interaction. This work addresses challenges which arise in the context of recognizing arbitrary human actions from body-worn sensors. Chiefly the invariance to temporal scaling of events, coping with unlabeled data and estimating an appropriate model complexity. In order to deal with the severe case of unlabeled data, a method is proposed based on dynamic time alignment of Gaussian mixture model clusters for matching actions in an unsupervised temporal segmentation. In facilitation of this, an extensive corpus of continuous motion sequences composed of everyday tasks was recorded as analysis scenarios. The technique achieved an average accuracy of 72% for correctly merging actions performed by different participants. With labeled data and recognition models designed for particular classes, an accuracy of 89% was achieved in classifying the motion of participants left out of the modeling process. These results are contrasted with benchmark methods for recognition in a systematic validation revealing, in particular, an improved performance for mixture model prediction utilizing segments.

1. Introduction

Algorithms for the recognition and synthesis of human motion have important applications in areas such as computer graphics, surveillance, health care and robotics. For health care, motion involving a gait, manual handling or hand-eye coordination can reveal the progression of fatigue or debilitating conditions [1]. Additionally, systems for tracking the daily activity level [2] or falls in elderly patients [3] are valuable tools in health monitoring. Synthesizing models can generate realistic and arbitrarily long computer animations from a set of motion capture clips [4]. Context-aware manufacturing systems can display relevant information for a complex task based upon a recognized sequence of actions performed by a worker [5]. Furthermore, gesture recognition is crucial for human–robot communication cues and to control the trajectory of a robot [6]. With the use of motion sensors in this expanding array of applications, wearable inertial sensors in particular, are increasingly adopted due to improvements in miniaturization and wireless communications [7].

Given this context, consider the problem of recognizing human actions from a continuous stream of sensor data. With incoming data and a set of patterns of interest or classes, each new array of data points is assigned membership to the relevant class based upon a model of the mapping between previous data and class pairs. In practice, when recording long data sequences, a number of issues naturally arise. Firstly, any recorded data is initially lacking an example set of class labels. Since it is usually a time-intensive task to manually assign labels it is common to apply a number of exploratory clustering and temporal segmentation methods to elucidate a plausible structure to the data. Partitioning the data into meaningful subsequences based upon the detection of change-points can significantly reduce the work required in attaching these labels [8,9]. Secondly, if an a priori set of class labels is associated with the data, this set may not sufficiently describe the range of significant patterns apparent to a human observer or additional patterns are superfluous for the application. This situation is often handled by introducing a null class to model irrelevant data. Lastly, given labeled examples of all relevant classes the remaining challenge is to compose a generalized description of an action originating from different recordings independent of discrepancies in kinematic structure, style or speed.

Temporal segmentation is certainly useful for processing data in real-time to assign these descriptions. However, if clustering similar segments, referred to here as temporal clustering, yields labels which align with those of a human observer, then the time required to amend the result is minimal. This is especially useful for applications where significant events are difficult to detect by inspection, due to a high dimensionality, and automating the discovery of such events may lead to a greater understanding of the data. Clearly, in cases such as human action recognition where the data is readily visualized, feedback to group particular events

* Corresponding author. Tel.: +61 4 34140414; fax: +61 2 42213236.
E-mail address: mf91@uowmail.edu.au (M. Field).

http://dx.doi.org/10.1016/j.patcog.2015.03.004
0031-3203/

will improve the ultimate recognition outcome. Quantifying the extent of this benefit and analyses of the bridge between event discovery and recognition are crucial to implementing these methods on a wider range of applications.

In this paper, an algorithm is presented for temporal segmentation leading to a method for the clustering of sequences of motion data for circumstances when no initial labels are available. The algorithm employs Gaussian mixture models (GMM) to represent human motion as a sequence of postures or motion primitives. Subsequences of primitives are identified through frequency analysis and compared via dynamic time warping in order to cluster similar sequences. The method is then extended to a motion recognition scenario to assess the accuracy of predictions as feedback on example class membership is available. In order to assess the performance extensive recordings of human motion data involving a set of everyday activities were collected from body-worn inertial sensors.

The remainder of this paper is organized as follows. Section 2 analyzes the related literature and details the scope of the paper. Section 3 describes the experimental equipment, the procedures of data capture and the features used for recognition. Section 4 explains the new approach of temporal segmentation and clustering while Section 5 presents the results according to a set of validation procedures. Section 6 discusses the issues associated with applications and the limitations of the approach which is then followed by concluding remarks.

2. Related work

Detecting time series segments is often an important initial phase in the pre-processing of collected data and is obviated by careful off-line preparation in many recognition problems. In this section, techniques for segmenting continuous data and related literature involving human motion recognition are reviewed before outlining the scope and contribution of this paper.

2.1. Temporal segmentation

Segments can be described by a set of time stamps indicating the beginning and ending point of an event. If possible, it is advantageous to use application specific sensors to detect these change points. Ward et al. [5] used the relative sound intensity between microphones positioned on the arm to indicate behavioral transitions in a work station. However, in the general case, segmentations must be deduced by directly analyzing the motion, which often involves the use of sliding window heuristics or time series modeling. There are a number of ways to detect changes between adjacent time windows. Many are centered around determining a descriptive model of the data which permits a distance measure between a pair of data sequences [8,10]. Barbic et al. proposed to use the Mahalanobis distance between a Gaussian distribution of a specific time frame and the subsequent sample to detect a significant change. Of the methods tested a distribution based on Probabilistic Principal Component Analysis (PPCA) features was claimed to have achieved the highest precision and recall for detection within a specified time bound. In [11] a dynamic hidden Markov model (HMM) was used to describe a window of time, and segment points were detected from exceeding a cost factor in describing new data points as unique states. This approach was also extended to segmentations of human motion capture [12].

An alternative to finding segmentations locally is to model the entire time series and detect patterns within the predicted state space. Sampling models of hierarchical HMMs can organize the data into proposal classes and, as a consequence, segments [13], however significant training time is required for long high dimensional data sets. Building the model incrementally may alleviate the burden of computational complexity while still clustering the dynamics [14]. On the contrary, some recent work has investigated clustering the data independent of time and analyzing subsequences of states with dynamic programming. For instance, Zhou et al. [15] used a set of k-means clusters to model the data points and developed a dynamic programming algorithm to cluster related segments and fine-tune their respective lengths. This approach compared favorably in accuracy with the sampled HMM-based methods. The segmentation of motion data with a GMM has been investigated in the literature [8,16] but in this paper, the notion is extended beyond the detection of change-points and towards recognizing the identity of repeated segments.

It is important to note that definitions of candidate segments in human motion vary among the applied literature. Some studies are concerned with segmenting short, limb specific actions such as ‘lift a leg’, ‘raise an arm’ or ‘squat’ [12], while others segment broad classes of actions, such as ‘walking’ or ‘boxing’, which are generally composed of a combination or sequence of simpler actions [8,17]. The actions considered further in this paper are of the more challenging latter variety.

2.2. Motion recognition

Human action recognition has had significant attention in the analysis of video streams [4], concentrating particularly on feature extraction rather than segmentation and recognition. Although feature extraction is less problematic for analyzing data from wearable sensors, in previous research the data is seldom modeled due to the high computational cost involved in unsupervised learning. The data is therefore commonly organized in a database structure in a reduced feature space leading to a database query system [18]. The methods pursuing this purpose usually exploit key frame techniques and a nearest neighbor distance metric to return similar frames and predict a class membership [19]. These distance measures are often affected by time scaling but methods have been proposed to deal with local scaling using dynamic time warping (DTW) [20], or a uniform scaling of candidate sequences [21]. The disadvantages of these methods involve the reliance on an existing categorized database and the cost of retaining and searching the original data.

Rather than rely upon large databases, modeling the data as a continuous process can lead to efficient descriptions of the data [22]. Work by Ward et al. [5] combined a pair of inertial sensors with microphones to classify actions in a staged workshop setting using HMMs with a manually crafted number of states. Their work also included extensive validation leaving out either repeated recordings or entire data sets from each of the 5 participants for testing. Gyllensten et al. [23] used a single accelerometer attached to the pelvis to recognize everyday actions with an ensemble classifier. Expanding upon the sensing capability, Altun et al. [24] used a set of 5 inertial sensors to discriminate between 19 activities performed by 8 subjects. In the leave-one-out test, the Support Vector Machines (SVM) and k-Nearest-Neighbors (k-NN) achieved 88% and 87% detection accuracy respectively. Even larger public databases, albeit from a limited set of wearable inertial sensors, of recorded daily activities [25], have been classified with a semi-supervised SVM to accuracies of approximately 76% [2]. Arm-mounted sensors have also been used to segment and classify realistic daily actions with HMMs from 4 subjects [26].

Although these findings demonstrate the practical use of inertial sensors in action recognition, the methodologies and features used typically result in decision thresholds rather than an encoding or representation of the motion. This paper adapts the latter strategy. Analyzing the motion, as composed of a sequence

Please cite this article as: M. Field, et al., Recognizing human motions through mixture modeling of inertial data, Pattern Recognition (2015), http://dx.doi.org/10.1016/j.patcog.2015.03.004