Pattern Recognition 48 (2015) 568–579

Contents lists available at ScienceDirect

Pattern Recognition

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

Trajectory-based human action segmentation

Luís Santos [a,1], Kamrad Khoshhal [a,2], Jorge Dias [a,b,*]

[a] University of Coimbra, Portugal
[b] Khalifa University, United Arab Emirates

ARTICLE INFO

Article history:
Received 26 June 2013
Received in revised form
7 July 2014
Accepted 17 August 2014
Available online 27 August 2014

Keywords:
Motion segmentation
Classification framework
Signal processing
Motion variability
Adaptive sliding window

ABSTRACT

This paper proposes a sliding window approach, whose length and time shift are dynamically adaptable in order to improve model confidence, speed and segmentation accuracy in human action sequences. Activity recognition is the process of inferring an action class from a set of observations acquired by sensors. We address the temporal segmentation problem of body part trajectories in Cartesian Space in which features are generated using Discrete Fast Fourier Transform (DFFT) and Power Spectrum (PS). We pose this as an entropy minimization problem. Using entropy from the classifier output as a feedback parameter, we continuously adjust the two key parameters in a sliding window approach, to maximize the model confidence at every step. The proposed classifier is a Dynamic Bayesian Network (DBN) model where classes are estimated using Bayesian inference. We compare our approach with our previously developed fixed window method. Experiments show that our method accurately recognizes and segments activities, with improved model confidence and faster convergence times, exhibiting anticipatory capabilities. Our work demonstrates that entropy feedback mitigates variability problems, and our method is applicable in research areas where action segmentation and classification is used. A working demo source code is provided online for academical dissemination purposes, by requesting the authors.

1. Introduction

Action recognition is an active research topic within the scientific community, with several applications, which include human–machine interfaces, intelligent video surveillance, video indexing and analysis. The action segmentation problem is a key issue in action recognition and may be divided into two stages: (1) learning and (2) classification. The learning stage often involves a data preprocessing step to find alternative, discriminant representations for different properties of the input signal. In this work, we consider a data driven probabilistic representation for the action model, which is learned from a set of training data. This action model is posteriorly used to identify to which action class each observable feature belongs to.

A popular applied method to this problem is the sliding window approach. The window is used to progress sequentially through the input signal, creating data segments from which features are extracted. This method is popular because of its direct integration with the majority of classification algorithms. However, fixed parameter values are a significant cause of classifier under-performance: slow convergence and/or borderline decisions (e.g. [1]). Choosing the ideal parameter values is not a trivial task and an optimal selection may differ for different performers and/or actions. Thus in this paper, we present a dynamically adaptive sliding window, where classification entropy is used to adjust the window length and time shift parameters at every step.

1.1. Action segmentation issues

The execution of actions differs from person to person. Factors like rigidly defined performance instructions, mobility restrictions introduced by the experimental set-up, cultural or anatomical characteristics are known to introduce variability. The majority of action models usually rely on a set of assumptions, which interfere with the classification of live executions and present some challenges. In our work, we are addressing the following problems:

- Frameworks can present high classification accuracy and the majority of the correct decisions are of low confidence. This is specially true as the number of different actions grows.

* Corresponding author. Tel.: +351 239 796 219, +351 918 711 531, +351 239 796 303; fax: +351 239 406 672.
E-mail addresses: luis@isr.uc.pt (L. Santos), kamrad@isr.uc.pt (K. Khoshhal), jorge@isr.uc.pt, jorge.dias@kustar.ac.ae (J. Dias).
URLS: http://www.isr.uc.pt/~luis (L. Santos), http://www.isr.uc.pt/~kamrad (K. Khoshhal), http://www.deec.uc.pt/~jorge (J. Dias).
[1] Tel.: +351 919 656 142; fax: +351 239 406 672.
[2] Tel.: +351 910 200 750; fax: +351 239 406 672.

http://dx.doi.org/10.1016/j.patcog.2014.08.015
0031-3203/

- The time it takes for a model to make a decision is highly dependent on the generated features, whereas being able to anticipate a decision is an issue of interest for an accurate temporal segmentation.

Approaches within action segmentation somehow try to address these factors. In this research, we are focused on extending our previous work using a fixed length sliding window approach [2,3], improving our segmentation solution to cope with classification performance issues. A survey on action segmentation [4] identifies other works which also use fixed length sliding windows [5–8]. In some of these works, the classification framework is augmented with multiple concurrent classifiers using windows of different lengths at the expense of increasing computational cost. Supported by examples in the literature, the following paragraphs summarize the main key problems in fixed parameter sliding window approaches.

A sliding window approach with fixed parameters is used in [9] to detect events in long video sequences. They analysed the delay (measured in frames) between ground truth annotations and the output of a classifier using the following parameters: a window size of 64 frames and a 8 frame time shift. Since an event temporal duration is variable, the fixed sliding window caused sample misclassification. In [10], the size of the sliding window is given in seconds (4 s) and it was used to detect unusual activities in video sequences. Result analysis shows that segmentation is not perfect and the reason for such large window size was to make sure that the buffer had enough signal information. Consequently, these large data samples contained higher rates of outlier information, which increase the number of borderline decisions. In [11] a sliding window was tested with two different sized, 48 and 24, frames. These were applied to video segmentation in the classification of human actions. Experimental results were presented without including classification decisions which contain transition from one action to another. Despite the application of this strategy, excluding transition frames did not prevent segment misclassification.

In other works, sliding window approaches are integrated with other techniques. For example, they can be integrated with Dynamic Time Warping [12,13], or Grammars [14,15]. However, methods that allow to dynamically adjusting the sliding window parameters in action segmentation are rarely explored. In [16], the window parameters are adjustable from sensor based events and dependent on the signal processing techniques. However, authors conclude that their approach is restricted by the application of the selected algorithms and sensors. In [17], a new type of self-adaptive sliding window is proposed for data mining. The parameters are adjustable based on the signal properties. While results seem to be satisfactory, the success of the proposed technique depends on the existence of specific signal properties. We were not able to find in the literature sliding approaches with dynamic parameters that are independent of the type of signal properties or processing algorithms.

1.2. Other works related on action segmentation

A recent survey by Weinland et al. [4] has identified three major action segmentation categories: sliding window, boundary detection and grammar concatenation. The already reviewed *sliding windows* are used to divide a motion sequence into multiple overlapping segments, which are bounded by the window limits. The information within the window may or may not be processed for alternative representations. Each candidate segment (or equivalent representation) is then used for sequential classification. The success of this approach strongly depends on the discriminant abilities of the generated representations. As mentioned this technique is easily integrated with the majority of static and dynamic classifiers. The major drawbacks of this technique are computational burden, and the need of multiple window sizes to overcome the variability problem. *Boundary detection* methods generally identify discontinuities or local extrema in observed motion signals. The boundaries usually define an implicit basic action taxonomy, without however depending on specific class definitions. A branch of works identifies boundary at the cost of the dynamics of the observed signal, such as [18,19]. Others depend on geometric property changes observed through techniques like Principal Component Analysis [20] or piecewise arc fitting models [21,22]. A related research addresses the segmentation problem from the subspace separation perspective, exploring the so-called Agglomerative Lossy Compression [23]. In [24], the authors apply Singular Value Decomposition (SVD) to a long sequence of optical flow images in order to detect trajectories' discontinuities within SVD component trajectories. Ogale et al. [25] also explore optical flow of body silhouettes, performing segmentation by detecting minima and maxima values of the absolute value sequence. A method using features from visual hulls is developed in [26]. This category of approaches is very sensitive to noise and other related errors (e.g. camera perspectives). Additionally, it allows generic segmentation, but is not particularly suitable for labelling purposes. The focus is on boundary identification rather than interpretation of intermediate data. Lastly, Weinland et al. [4] identify *Grammars* as another category. The common approach is to model state transitions between actions, where Hidden Markov Models (HMM) are a popular approach. Multiple methods can be used to generate features. Some examples are curvature scale space and centroid distance function [27], joint angles alone [28,29], or together with velocity profiles [30], dynamic system representations [31–33] and geometrical property encoding [34]. These are applied to segment and label action sequences at the expense of computing a minimum-cost path through the model using techniques like Viterbi path, Conditional Random Fields or Markov Models. However, these methods rely on the comprehensiveness of state grammars, which may jeopardize the model effectiveness and the generalization purpose, if large amount of training data is not available.

We can say that temporal action segmentation is implicitly addressed in most problems of action classification at some point of their research. The majority of research is done in computer vision and applied to image sequences, where each frame is classified consequently generating a temporal sequence of associated action labels, such as in [35,36]. More classical vision-based approaches only consider data from the current image frame, attempting to find a class that represents the acquired data more closely. There are in fact other works that consider collections of multiple images, as it happens in a sliding window paradigm. But again, these also use a pre-defined number of images and time shifts (e.g. [37]).

1.3. Definitions and problem statement

A motion instance is defined as a contiguous sequence of human body movements, which is composed of a concatenation of different actions. Let motion instance Ω be a sequence of 3-D Cartesian coordinates Y, defining a discrete trajectory of random duration T (measured in *frames*), for a body part such that

$$\Omega = \begin{bmatrix} Y_1 \\ \vdots \\ Y_T \end{bmatrix}, \quad Y \in \mathbb{R}^3 \text{ and } T \in \mathbb{N} \tag{1}$$