Segmentation of Action Streams
Human Observers vs. Bayesian Binning

Dominik Endres, Andrea Christensen, Lars Omlor, and Martin A. Giese

Section for Computational Sensomotorics, Dept. of Cognitive Neurology,
University Clinic, CIN, HIH and University of Tübingen,
Frondsbergstr 23, 72070 Tübingen, Germany
dominik.endres@klinikum.uni-tuebingen.de
{andrea.christensen,martin.giese}@uni-tuebingen.de

Abstract. Natural body movements are temporal sequences of individual actions. In order to realise a visual analysis of these actions, the human visual system must accomplish a temporal segmentation of action sequences. We attempt to reproduce human temporal segmentations with Bayesian binning (BB)[8]. Such a reproduction would not only help our understanding of human visual processing, but would also have numerous potential applications in computer vision and animation. BB has the advantage that the observation model can be easily exchanged. Moreover, being an exact Bayesian method, BB allows for the automatic determination of the number and positions of segmentation points. We report our experiments with polynomial (in time) observation models on joint angle data obtained by motion capture. To obtain human segmentation points, we generated videos by animating sequences from the motion capture data. Human segmentation was then assessed by an interactive adjustment paradigm, where participants had to indicate segmentation points by selection of the relevant frames. We find that observation models with polynomial order ≥ 3 can match human segmentations closely.

1 Introduction

Temporally segmenting (human) action streams is interesting for a variety of reasons: firstly, if we had a model which reproduced human segmentations closely, it might reveal important insights in human action representation. Previous work in this direction has studied in detail the segmentation of sequences of piecewise linear movements in the two-dimensional plane [23,1]. Secondly, a good temporal segmentation would have numerous applications in the field of computer vision. Worth mentioning in this context is the Human Motion Analysis (HMA) [24]. HMA concerns the detection, tracking and recognition of people from image sequences involving humans and finds its application in many areas such as smart surveillance and man-machine interfaces. Thirdly, extraction of important key frames by improved motion segmentation would not only contribute to computer vision research but also to computer graphics and motion synthesis. Animations of human motion data can be done with less computational costs if the key frames are defined optimally (e.g. [6,5]).

J. Bach and S. Edelkamp (Eds.): KI 2011, LNAI 7006, pp. 75–86, 2011.

While most researchers base their temporal segmentation approaches on real video data and focus on the computer vision problem to analyse human motion data by tracking of skeleton models or feature sequences [21,2,13], we address here specifically the problem of the segmentation of action streams based on motion capture data. We compare Bayesian binning (BB) for segmentation of human full-body movement with human responses, which were assessed in an interactive video segmentation paradigm.

BB is a method for modelling data with a totally ordered structure, e.g. time series, by piecewise defined functions. Its advantages include automatic complexity control, which translates into automatic determination of the number and length of the segments in our context. BB was originally developed for density estimation of neural data and their subsequent information theoretic evaluation [8]. It was later generalised for regression of piecewise constant functions [14] and further applications in neural data analysis [10,9]. Concurrently, a closely related formalism for dealing with multiple change point problems was developed in [11].

We give a concise description of the data recordings in section 2, since these data have not been published before. The psychophysical experiments and their results are described in section 3. We use BB for the segmentation of joint angle data obtained by motion capture in section 4. Furthermore, we show how to use BB with non-constant observations models in the bins. In section 5 we present the segmentations achieved by BB and compare them with the psychophysical results. Finally, we discuss the advantages and limitations of our approach and give an outlook for further investigations in section 6.

2 Kinematical Data

The action streams we studied are solo Taekwondo activities performed by ten internationally successful martial artists. Each combatant performed the same fixed sequence of 27 kicks and punches, forming a so called *hyeong*. A complete hyeong had a full length of about 40 seconds. The kinematical data was obtained by motion capture using a VICON 612 system with 11 cameras, obtaining the 3D positions of 41 passively reflecting markers attached to the combatants' joints and limbs with a 3D reconstruction error of below 1 mm and at a sampling frequency 120 Hz.

The use of the obtained kinematical data was twofold. First, joint angle trajectories were computed from a hierarchical kinematic body model (skeleton) which was fitted to the original 3D marker positions. The rotations between adjacent segments of this skeleton were described by Euler angles, defining flexion, abduction and rotations about the connecting joint (e.g. [19,22]). Second, from the derived joint angle trajectories we created movie clips showing computer animations of the Taekwondo movements. Those videos served as stimuli in our psychophysical experiment to obtain human segmentation ratings.