Emulating human observers with Bayesian binning: segmentation of action streams

DOMINIK ENDRES, ANDREA CHRISTENSEN, LARS OMLOR, MARTIN A. GIESE, Section for Theoretical Sensomotorics, Department of Cognitive Neurology, University Clinic Tübingen, University of Tübingen and Hertie Institute for Clinical Brain Research and Center for Integrative Neuroscience

Natural body movements arise in the form of temporal sequences of individual actions. During visual action analysis, the human visual system must accomplish a temporal segmentation of the action stream into individual actions. Such temporal segmentation is also essential to build hierarchical models for action synthesis in computer animation. Ideally, such segmentations should be computed automatically in an unsupervised manner. We present an unsupervised segmentation algorithm that is based on Bayesian binning (BB) and compare it to human segmentations derived from psychophysical data. 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 applied this method to motion capture sequences from martial arts and compared the results to segmentations provided by humans from movies that showed characters that were animated with 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. Results show a good agreement between automatically generated segmentations and human performance when the trajectory segments between the transition points were modelled by polynomials of at least third order. This result is consistent with theories about differential invariants of human movements.

Categories and Subject Descriptors: I.2.10 [**Artificial Intelligence**]: Vision and Scene Understanding—*Motion*

Additional Key Words and Phrases: motion capture, action segmentation, unsupervised learning, Bayesian methods

Cite as:
Endres D., Christensen C., Omlor L. and Giese M.A. (2011). Emulating human observers with Bayesian binning: segmentation of action streams. *ACM Transactions on Applied Perception (TAP)*, **8**(3), 16:1-12.

Author's address: Section for Computational Sensomotorics, Department of Cognitive Neurology, University Clinic Tübingen, University of Tübingen and Hertie Institute for Clinical Brain Research and Center for Integrative Neuroscience, Frondsbergstr. 23, 72070 Tübingen, Germany.
✉: dominik.endres@klinikum.uni-tuebingen.de, andrea.christensen@uni-tuebingen.de, martin.giese@uni-tuebingen.de

DOI 10.1145/0000000.0000000 http://doi.acm.org/10.1145/0000000.0000000

1. INTRODUCTION

The temporal segmentation of human action streams is interesting for several reasons: firstly, good models of human segmentation performance might reveal important insights into the structure of action representations in the brain. With this goal, previous work has studied in detail the segmentation of sequences of piecewise linear movements in the two-dimensional plane [Shipley et al. 2004; Agam and Sekuler 2008]. Secondly, the automatic determination of good segmentation points is essential for many learning-based methods for movement synthesis, which approximate individual movements by parametric models for individual segments that are obtained by interpolation from example trajectories. Examples are models based on PCA [Safonova et al. 2004; Ilg et al. 2004] or the classical verb-adverbs approach [Rose et al. 1998] that rely on accurately pre-segmented data. Likewise, motion graphs or approaches that synthesise novel sequences by pasting together fragments from a motion capture data base that fulfil additional constraints [Arikan and Forsyth 2002; Kovar et al. 2002] ideally require segment boundaries that are consistent with the parsing in human perception. While classically such segmentations have been generated by hand, this approach becomes infeasible for larger sets of motion capture data. This makes automatic segmentation an essential problem for learning-based computer animation.

We therefore address the problem of the temporal segmentation of action streams represented by motion capture data. We compare Bayesian binning (BB)[Endres and Földiák 2005] for segmentation of human full-body movement with human responses, which were assessed in an interactive video segmentation paradigm. BB is an approach to model data with a totally ordered structure, such as time series, by functions which are defined piecewise. The method can determine automatically the appropriate number and length of temporal segments via Bayesian model selection. BB was originally developed for density estimation of neural data [Endres and Földiák 2005]. It was later generalised for regression of piecewise constant functions [Hutter 2007] and further applications in computational neuroscience [Endres and Oram 2010]. Concurrently, a closely related formalism for dealing with multiple change point problems was developed in [Fearnhead 2006].

This paper is structured as follows: We first describe the data recordings in section 2. The psychophysical experiments and their results are presented in section 3. We then describe the application of BB for the segmentation of joint angle data in section 4 and extend BB for non-constant observation models in the bins. In section 5 we demonstrate the results achieved by BB and compare them with the segmentations from the psychophysical experiments. Finally, advantages and limitations of our approach are discussed in section 6.

2. MOVEMENT RECORDINGS

This study is based on motion capture data recorded from ten internationally successful martial artists performing the same solo taekwondo pattern (*hyeong*). Each hyeong consists of a series of 27 kicks and punches linked together in a fixed sequence of approximately 40 seconds length. Motion capture data was recorded using a VICON 612 motion capture system with 11 cameras, obtaining the 3D positions of 41 passively reflecting markers attached to the combatants' joints and limbs. The algorithmic segmentation is based on joint angle trajectories which were computed from an hierarchical kinematic body model (skeleton) fitted to the original 3D marker positions. The rotations between adjacent segments of this skeleton were first described by Euler angles, defining flexion, abduction and rotations about the connecting joint and were transformed to an axis angle representation (e.g. [Roether et al. 2009]).

3. PSYCHOPHYSICAL EXPERIMENT

As reference for the segmentation performance of our algorithmic approach we conducted a psychophysical study. In this experiment human observers segmented video clips showing the same taekwondo movements animated as volumetric puppets.

3.1 Stimuli

The recorded motion capture data of each hyeong was split manually into five sub-sequences of comparable length each containing between three and eight separable taekwondo moves. Animations of those sub-sequences of movements using a custom-built volumetric puppet served as stimuli in our experiments. An illustration of the stimulus material can be found in the supplementary information, snapshots of the animated movements are shown in fig. (1)A. Five of the overall 50 animated videos, corresponding to the complete hyeong of one combatant, served as training stimuli to familiarise the participants with the videos as well as with the experimental procedure. In order to limit the duration of the experiment and to prevent fatigue of the participants, only 25 animated sub-sequences from the hyeongs of five representative combatants were used in the final experiment. Each animation was repeated three times resulting in 75 segmentation trials per subject. Stimuli subtended approximately 4 x 8.6 degrees of visual angle and were presented on a computer screen viewed from a distance of 50 cm.

3.2 Participants

Thirteen participants (mean age 26 years 6 month, ranging from 21 years 11 month to 38 years 11 month, 10 female) segmented the animated movies in our experiment. All had normal or corrected-to-normal vision, gave informed written consent and were paid for their participation.

3.3 Segmentation task

The segmentation task was identical during the training and the test phase: In every trial the movie to segment was first presented twice to enable the subjects to familiarise themselves with the stimulus. While showing the animation for the third time the participants segmented the movements by pressing a key at any point which they judged as the endpoint of one single, separable movement. Afterwards, participants had the opportunity to watch their own segmentation of the current movie and to correct themselves up to two times if they were not satisfied with the result. However, what exactly defines *one single movement* and the corresponding *endpoint* was left to the own judgement of the participants, and no feedback was given at any time during the training or the testing.

3.4 Segmentation results of human observers

The motion segmentations for the complete hyeong of one representative taekwondo combatant as indicated by the human observers are shown in fig. (1)B. The 39 rows of black dots correspond to the three segmentations of each of the thirteen participants. Each single black dot represents the perception of one endpoint of a movement within the sequence of movements. The red lines indicate video boundaries between the 5 sub-sequences the complete hyeong was split into. Since neither a definition of an endpoint of one movement nor any feedback was given the interpretation of one separable movement differed somehow between subjects. While most subjects tended to segment the videos on a very fine-grained level with many endpoints, two subjects concentrated on the coarse separation of the hyeong by setting only 5 respectively 7 segmentation points. Nevertheless the defined endpoints are still very consistent across subjects and the overall mean number of perceived movement endings (22.08 (standard error 3.01)) is close to the actual number of endings of individual taekwondo techniques (27). Comparing the position of perceived and expected endpoints, subjects had a mean hit rate of 0.69 (standard error 0.05) within an accuracy window of less than ± 250 ms. These results are in accordance with previous findings about the agreement of human raters on boundary placing in movement sequences [Dickman 1963; Newtson and Engquist 1976; Zacks et al. 2009].

4. BAYESIAN BINNING FOR ACTION SEGMENTATION

We now give a brief specification of the BB model which we will use to segment joint angle data and to model human key press data.

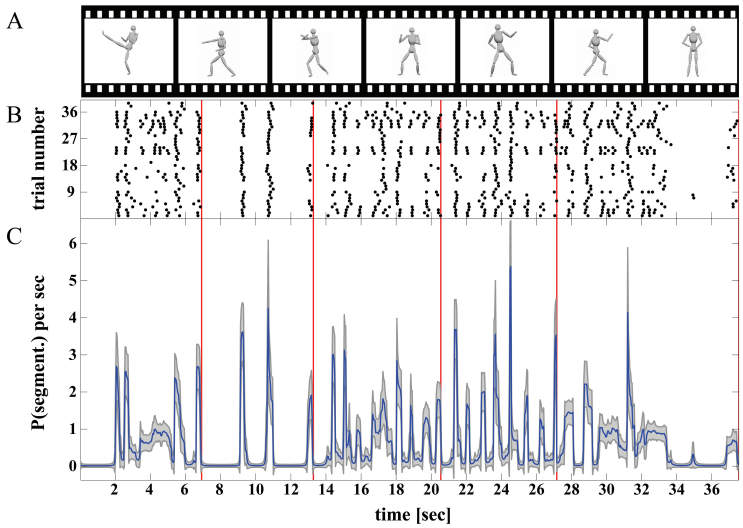

Fig. 1. Human Motion Segmentation. A) Illustration of Stimuli. Snapshots taken from the stimuli videos showing the custom-built volumetric grey puppet performing different taekwondo kicks and punches. B) Segmentation points over time as marked by 13 participants. Each black dot corresponds to one approved key press, indicating the perception of a transition between two taekwondo movements. Red lines show video boundaries between the 5 sub-sequences. C) Predictive segmentation density estimated from key presses. Estimation was carried out by Bayesian binning with a Bernoulli-Beta observation model (see section 4). Blue line represents the predictive segmentation using Bayesian binning, the shaded grey area indicates the probability ± one std.dev. Red lines as in B.

4.1 Objective

Our objective is to model a time series D in the time interval $[t_{min}, t_{max}]$. We would like to be able to handle D corrupted by (large amounts of) noise, let the model complexity be driven by D, and draw conclusions (e.g. change point estimates) from small amounts of data. Thus we choose a Bayesian approach. We discretise $[t_{min}, t_{max}]$ into T contiguous intervals of duration $\Delta t = (t_{max} - t_{min})/T$. Choose Δt small enough to capture all relevant features of the data. We model the generative process of D by $M+1$ contiguous, non-overlapping bins, indexed by m and having upper boundaries $k_m \in \{k_m\}$. The bin m therefore contains the time interval $T_m = (\Delta t\ k_{m-1}, \Delta t\ k_m]$. Let D_m be that part of the data which falls into bin m. We assume that the probability of D given $\{k_m\}$ can be factorised as

$$P(D|\{k_m\}, M) = \prod_{m=0}^{M} P(D_m|k_{m-1}, k_m, M) \tag{1}$$

4.2 Prior on $\{k_m\}$

Assuming we have no preferences for any bin boundary configuration (other than $m' < m \Rightarrow k_{m'} < k_m$), the prior is just the reciprocal of the number of possibilities in which M ordered bin boundaries can be distributed across $T-1$ places, i.e.

$$P(\{k_m\}|M) = \left(\begin{array}{c} T-1 \\ M \end{array} \right)^{-1}. \tag{2}$$

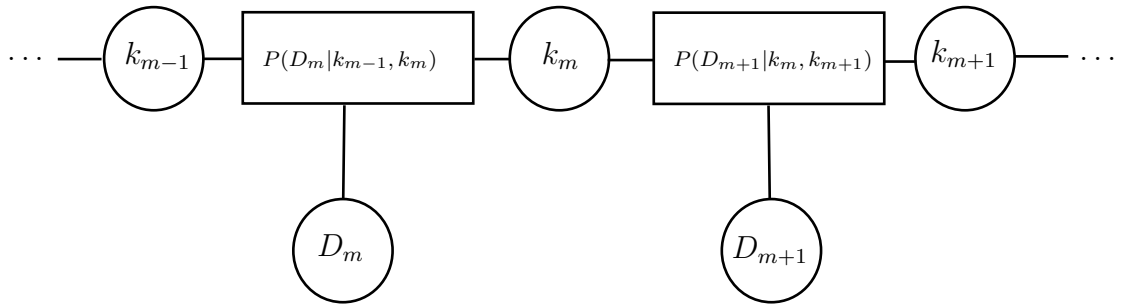

Fig. 2. Fragment of a factor graph for Bayesian binning. Each round node represents a random variable, here: bin boundaries $\{k_m\}$ and observable joint angle data D_m. A rectangular node depicts a factor in the likelihood eqn.(1), connected to those variables which appear in it. The constant prior factor (eqn.(2)) has been omitted since it has no influence on the connectivity structure of the graph. Note that each data node D_m is connected to one factor node $P(D_m|k_{m-1}, k_m)$ only (for each possible configuration of the $\{k_m\}$). Thus, the graph is singly connected and exact marginals can be computed with the sum-product algorithm efficiently. For details, see text.

4.3 Prior on M

Assuming we have no preference for any model complexity (i.e. number of bin boundaries), let $P(M) = \frac{1}{T}$ since the number of bin boundaries M must be $0 \leq M \leq T - 1$.

4.4 Evaluating the posterior of $\{k_m\}$

In the context of temporal segmentation, the most relevant posterior is that of the $\{k_m\}$ for a given M:

$$P(\{k_m\}|D, M) = \frac{P(D|\{k_m\}, M)P(\{k_m\}|M)}{P(D|M)}. \tag{3}$$

The evaluation of the denominator, $P(D|M)$, naïvely requires a computational effort of $\mathcal{O}(T^M)$, because each of the M bin boundaries can be in $\approx T$ many places. However, it is possible to reduce this effort to $\mathcal{O}(MT^2)$, as described in [Endres and Földiák 2005]: $P(D|M)$ can be evaluated exactly with an instance of the *sum-product* algorithm [Kschischang et al. 2001], because the factor graph corresponding to BB is singly connected (see fig. (2)). The reason for this single-connectedness can be understood by inspection of eqn.(1): the factors $P(D_m|k_{m-1}, k_m, M)$ depend on non-overlapping parts of the data (for a given configuration of the $\{k_m\}$). Hence, every data node D_m in fig. (2) is only connected to one data node.

4.5 Observation model $P(D|\{k_m\})$ for human segmentation events

Human segmentations (i.e. key presses by observers) are binary events. We therefore use a Bernoulli process with a conjugate Beta prior (one per bin) for these data. A conjugate prior allows for the evaluation of expectations and marginal probabilities in closed form. This is analogous to modelling neural spike trains with BB [Endres and Oram 2010]. Thus, for a segmentation event $e(t) \in D$ at time t in bin m, i.e. $t \in T_m$ we have

$$P(e(t)|t \in T_m) = P_m \tag{4}$$

$$p(P_m) = \mathrm{B}(P_m; \gamma_m, \delta_m) \tag{5}$$

where $\mathrm{B}(P_m; \gamma_m, \delta_m)$ is the Beta density with parameters γ_m, δ_m (see e.g. [Bishop 2007]).

4.6 Observation model $P(D|\{k_m\})$ for joint angles

To construct an observation model for joint angles, we will assume independence across bins. We do not claim that this assumption is strictly fulfilled, but it yields good results on our data (see section 5). Hence, we only need to derive an observation model for one bin, and we drop the bin index m in the following. Moreover, assume that all time indexes t are relative to the lower boundary of the bin currently under consideration. Since joint angles are real numbers in $[-\pi, \pi)$, we could therefore use a multivariate von-Mises density or generalisations thereof [Marida and Jupp 2000]. However, for analytical convenience, particularly because the conjugate priors are tractable, we shall instead model joint angles with a multivariate Gaussian whose mean μ has a polynomial time dependence. An conjugate prior on the mean μ and the precision matrix $\mathbf{P}$ (inverse covariance) can then be derived via the exponential family construction (see e.g. [Bishop 2007]) and has an extended Gauss-Wishart density. We outline the derivation in some detail here, since it has not been published before.

Let $\vec{x}_t \in D$ be a L-dimensional column vector of joint angles at time t, and $S-1$ be the chosen polynomial order. Then

$$
\begin{aligned}
p(\vec{x}_t|\vec{\mu}, \mathbf{P}) &= \mathcal{N}(\vec{X}(t); \vec{\mu}, \mathbf{P}^{-1}) && (6)\\
p(\mathbf{P}|\nu, \mathbf{V}) &= \mathcal{W}(\mathbf{P}; \nu, \mathbf{V}) && (7)\\
\vec{\mu} &= \mathbf{a}\,\vec{t} && (8)\\
\vec{t} &= (t^0, t^1, \ldots, t^{S-1})^T && (9)
\end{aligned}
$$

Row l of the $L \times S$ matrix $\mathbf{a}$ contains the polynomial coefficients for joint angle l. Thus, the data likelihood (eqn.(6)) becomes

$$
\begin{aligned}
p(\vec{x}_t|\mathbf{a}, \mathbf{P}) &= \frac{\sqrt{|P|}}{\sqrt{2\pi}^L} \exp\left(-0.5(\vec{x}_t - \mathbf{a}\vec{t})^T\mathbf{P}(\vec{x}_t - \mathbf{a}\vec{t})\right) && (10)\\
&= \frac{\sqrt{|P|}}{\sqrt{2\pi}^L} \exp\left(-0.5\ \mathbf{tr}\left[(\vec{x}_t - \mathbf{a}\vec{t})(\vec{x}_t - \mathbf{a}\vec{t})^T\mathbf{P}\right]\right) && (11)
\end{aligned}
$$

Now suppose we observe a data set D of N joint angle vectors, $D = (\vec{x}_{t_0}, \ldots, \vec{x}_{t_{N-1}})$. We assume that there is no dependency between the $\vec{x}_{t_i}$beyond the polynomial time dependence of the mean. Thus, setting $Z_d(\mathbf{P}) = \frac{\sqrt{2\pi}^{LN}}{\sqrt{|P|}^N}$, the likelihood of D is

$$
\begin{aligned}
p(D|\mathbf{a}, \mathbf{P}) &= Z_d(\mathbf{P})^{-1} \exp\left(-0.5\sum_{i=0}^{N}\ \mathbf{tr}\left[(\vec{x}_{t_i} - \mathbf{a}\,\vec{t}_i)(\vec{x}_{t_i} - \mathbf{a}\,\vec{t}_i)^T\mathbf{P}\right]\right) && (12)\\
&= Z_d(\mathbf{P}) \exp\left(-0.5\ \mathbf{tr}\left[\left(\sum_{i=0}^{N}\vec{x}_{t_i}\vec{x}_{t_i}^T - \sum_{i=0}^{N}\left(\vec{x}_{t_i}\vec{t}_i^T\ \mathbf{a}^T + \mathbf{a}\,\vec{t}_i\vec{x}_{t_i}^T\right) + \sum_{i=0}^{N}\mathbf{a}\,\vec{t}_i\vec{t}_i^T\ \mathbf{a}^T\right)\mathbf{P}\right]\right) && (13)
\end{aligned}
$$

A prior on $\mathbf{a}$ which is conjugate[1] to this likelihood and in the exponential family can be constructed by choosing a $L \times S$ matrix $\mathbf{A}$ (the prior means of $\mathbf{a}$) and a $S \times S$, symmetric and positive definite matrix $\mathbf{B}$ (the concentration parameters of $\mathbf{a}$), such that

$$
p(\mathbf{a}|\mathbf{A}, \mathbf{B}, \mathbf{P}) = Z_g(\mathbf{B}, \mathbf{P})^{-1} \exp\left(-0.5\ \mathbf{tr}\left[(\mathbf{a} - \mathbf{A})\ \mathbf{B}\ (\mathbf{a} - \mathbf{A})^T\ \mathbf{P}\right]\right) \quad (14)
$$

[1] a prior is said to be conjugate to a likelihood, if the resulting posterior has the same functional form as the prior, see e.g. [Bishop 2007] for an introduction. This reference also contains an overview of exponential family distributions.

$$Z_g(\mathbf{B},\mathbf{P}) \;=\; \int d\mathbf{a}\; \exp\left(-0.5\;\mathbf{tr}\left[(\mathbf{a}-\mathbf{A})\;\mathbf{B}\;(\mathbf{a}-\mathbf{A})^T\;\mathbf{P}\right]\right) \tag{15}$$

where $Z_g(\mathbf{B},\mathbf{P})$ is a normalisation constant. We will demonstrate below (eqn.(25)) that $Z_g(\mathbf{B},\mathbf{P})$ does not depend on $\mathbf{A}$. To show that this prior is indeed conjugate to eqn.(12), we need to show that the posterior of $\mathbf{a}$ given D

$$p(\mathbf{a}|\mathbf{A},\mathbf{B},\mathbf{P},D) = \frac{p(D|\mathbf{a},\mathbf{P})p(\mathbf{a}|\mathbf{A},\mathbf{B},\mathbf{P})}{p(D|\mathbf{A},\mathbf{B},\mathbf{P})}. \tag{16}$$

has the same functional form as the prior. Using eqn.(12) and eqn.(14), the numerator of the r.h.s can be written as

$$\begin{aligned} p(D|\mathbf{a},\mathbf{P})p(\mathbf{a}|\mathbf{A},\mathbf{B},\mathbf{P}) \;=\; & Z_d(\mathbf{P})^{-1}Z_g(\mathbf{B},\mathbf{P})^{-1}\times \\ & \times\;\; \exp\left(-0.5\;\mathbf{tr}\left[\left(\sum_{i=0}^{N}\vec{x}_{t_i}\vec{x}_{t_i}^T - \sum_{i=0}^{N}\left(\vec{x}_{t_i}\vec{t}_i^T\;\mathbf{a}^T + \mathbf{a}\;\vec{t}_i\vec{x}_{t_i}^T\right) + \sum_{i=0}^{N}\mathbf{a}\;\vec{t}_i\vec{t}_i^T\;\mathbf{a}^T\right)\mathbf{P}\right]\right) \\ & \times\;\; \exp\left(-0.5\;\mathbf{tr}\left[\left(\mathbf{a}\mathbf{B}\mathbf{a}^T - \left(\mathbf{A}\mathbf{B}\mathbf{a}^T + \mathbf{a}\mathbf{B}\mathbf{A}^T\right) + \mathbf{A}\mathbf{B}\mathbf{A}^T\right)\mathbf{P}\right]\right) \end{aligned} \tag{17}$$

We now rewrite this expression by collecting terms and introducing the *posterior parameters*

$$\hat{\mathbf{B}} \;:=\; \mathbf{B} + \sum_{i=0}^{N}\vec{t}_i\vec{t}_i^T \tag{18}$$

$$\hat{\mathbf{A}}^T \;:=\; \hat{\mathbf{B}}^{-1}\left(\mathbf{B}\mathbf{A}^T + \sum_{i=0}^{N}\vec{t}_i\vec{x}_{t_i}^T\right). \tag{19}$$

$$\tag{20}$$

Note that eqn.(19) implies $\hat{\mathbf{B}}\hat{\mathbf{A}}^T = \mathbf{B}\mathbf{A}^T + \sum_{i=0}^{N}\vec{t}_i\vec{x}_{t_i}^T$, and eqn.(18) ensures that $\hat{\mathbf{B}}$ is positive definite, if $\mathbf{B}$ is. Additionally, let

$$\mathbf{U} := \sum_{i=0}^{N}\vec{x}_{t_i}\vec{x}_{t_i}^T + \mathbf{A}\mathbf{B}\mathbf{A}^T - \hat{\mathbf{A}}\hat{\mathbf{B}}\hat{\mathbf{A}}^T \tag{21}$$

Then

$$\begin{aligned} p(D|\mathbf{a},\mathbf{P})p(\mathbf{a}|\mathbf{A},\mathbf{B},\mathbf{P}) \;=\; & Z_d(\mathbf{P})^{-1}Z_g(\mathbf{B},\mathbf{P})^{-1}\;\exp\left(-0.5\;\mathbf{tr}\left[\mathbf{U}\mathbf{P}\right]\right) \\ & \times\;\; \exp\left(-0.5\;\mathbf{tr}\left[\left((\mathbf{a}-\hat{\mathbf{A}})\hat{\mathbf{B}}(\mathbf{a}-\hat{\mathbf{A}})^T\right)\mathbf{P}\right]\right) \end{aligned} \tag{22}$$

The denominator of eqn.(14) can be obtained from this numerator by integrating out $\mathbf{a}$. Since the first three factors on the r.h.s. of eqn.(22) do not depend on $\mathbf{a}$, we finally obtain

$$p(\mathbf{a}|\mathbf{A},\mathbf{B},\mathbf{P},D) = \frac{\exp\left(-0.5\;\mathbf{tr}\left[\left((\mathbf{a}-\hat{\mathbf{A}})\hat{\mathbf{B}}(\mathbf{a}-\hat{\mathbf{A}})^T\right)\mathbf{P}\right]\right)}{\int d\mathbf{a}\;\exp\left(-0.5\;\mathbf{tr}\left[\left((\mathbf{a}-\hat{\mathbf{A}})\hat{\mathbf{B}}(\mathbf{a}-\hat{\mathbf{A}})^T\right)\mathbf{P}\right]\right)}. \tag{23}$$

Identifying $Z_g(\hat{\mathbf{B}},\mathbf{P}) = \int d\mathbf{a}\;\exp\left(-0.5\;\mathbf{tr}\left[\left((\mathbf{a}-\hat{\mathbf{A}})\hat{\mathbf{B}}(\mathbf{a}-\hat{\mathbf{A}})^T\right)\mathbf{P}\right]\right)$ in eqn.(14), we see that the prior is conjugate.

For Bayesian binning, we need the marginal likelihood of the data in each bin (see eqn.(1)), which can be calculated by integrating eqn.(22) over $d\mathbf{a}$:

$$\begin{aligned} p(D|\mathbf{A},\mathbf{B},\mathbf{P}) &= \int d\mathbf{a}\, p(D|\mathbf{a},\mathbf{P}) p(\mathbf{a}|\mathbf{A},\mathbf{B},\mathbf{P}) \\ &= Z_d(\mathbf{P})^{-1} Z_g(\mathbf{B},\mathbf{P})^{-1} Z_g(\hat{\mathbf{B}},\hat{\mathbf{P}})\ \exp\left(-0.5\ \mathbf{tr}\left[\mathbf{UP}\right]\right). \end{aligned}$$

To evaluate this expression, we now compute the normalisation constant $Z_g(\mathbf{B},\mathbf{P})$. Let $\vec{u}$ be the $L \cdot S$-dimensional column vector obtained by stacking all columns of $\mathbf{a}-\mathbf{A}$. Denoting the Kronecker product with $\otimes$, it is straightforward to show that

$$\mathbf{tr}\left[\left((\mathbf{a}-\mathbf{A})\mathbf{B}(\mathbf{a}-\mathbf{A})^T\right)\mathbf{P}\right] = \vec{u}^T \mathbf{B}\otimes\mathbf{P}\vec{u} \tag{24}$$

by applying the identities 497 and 496 from the matrix cookbook [Petersen and Pedersen 2008]. Hence, the integral in $Z_g(\mathbf{B},\mathbf{P})$ is Gaussian with precision matrix $\mathbf{B}\otimes\mathbf{P}$:

$$Z_g(\mathbf{B},\mathbf{P}) = \int d\mathbf{a}\ \exp\left(-0.5\ \vec{u}^T(\mathbf{B}\otimes\mathbf{P})\vec{u}\right) = \frac{\sqrt{2\pi}^{LS}}{\sqrt{|\mathbf{B}|}^L \sqrt{|\mathbf{P}|}^S} \tag{25}$$

where we have used identity 491 from [Petersen and Pedersen 2008] for the determinant of a Kronecker product. Therefore, the marginal likelihood (for fixed precision matrix $\mathbf{P}$) is

$$p(D|\mathbf{A},\mathbf{B},\mathbf{P}) = \frac{\sqrt{|\mathbf{P}|}^N}{\sqrt{2\pi}^{LN}} \frac{\sqrt{|\mathbf{B}|}^L}{\sqrt{|\hat{\mathbf{B}}|}^L}\ \exp\left(-0.5\ \mathbf{tr}\left[\mathbf{UP}\right]\right) \tag{26}$$

Since we would also like to learn the precision matrix $\mathbf{P}$, we equip it with a Wishart prior, as stated above (eqn.(7)). The density of this prior is given by [Bishop 2007]:

$$p(\mathbf{P}|\nu,\mathbf{V}) = Z_w(\nu,\mathbf{V})^{-1} |\mathbf{P}|^{\frac{\nu-L-1}{2}}\ \exp\left(-0.5\ \mathbf{tr}\left[\mathbf{V}^{-1}\mathbf{P}\right]\right) \tag{27}$$

$$Z_w(\nu,\mathbf{V}) = 2^{\frac{\nu L}{2}} |\mathbf{V}|^{\frac{\nu}{2}} \Gamma_L\left(\frac{\nu}{2}\right) \tag{28}$$

$$\Gamma_L\left(\frac{\nu}{2}\right) = \pi^{\frac{L(L-1)}{4}} \prod_{j=1}^{L} \Gamma\left(\frac{\nu+1-j}{2}\right) \tag{29}$$

where $\Gamma(\cdot)$ is the gamma function.

We now show that this prior is conjugate to the marginal likelihood eqn.(26) by computing the posterior density of $\mathbf{P}$ given D:

$$p(\mathbf{P}|\nu,\mathbf{V},D) = \frac{p(D|\mathbf{A},\mathbf{B},\mathbf{P}) p(\mathbf{P}|\nu,\mathbf{V})}{p(D|\mathbf{A},\mathbf{B},\nu,\mathbf{V})} \tag{30}$$

After rearranging terms, the numerator of this posterior is

$$p(D|\mathbf{A},\mathbf{B},\mathbf{P}) p(\mathbf{P}|\nu,\mathbf{V}) = \frac{\sqrt{|\mathbf{B}|}^L}{\sqrt{|\hat{\mathbf{B}}|}^L} \sqrt{2\pi}^{-LN} Z_w(\nu,\mathbf{V})^{-1} |\mathbf{P}|^{\frac{N+\nu-L-1}{2}}\ \exp\left(-0.5\ \mathbf{tr}\left[(\mathbf{V}^{-1}+\mathbf{U})\mathbf{P}\right]\right) \tag{31}$$

We introduce the *posterior parameters*

$$\hat{\nu} := \nu + N \tag{32}$$

$$\hat{\mathbf{V}}^{-1} := \mathbf{V}^{-1} + \mathbf{U} \tag{33}$$

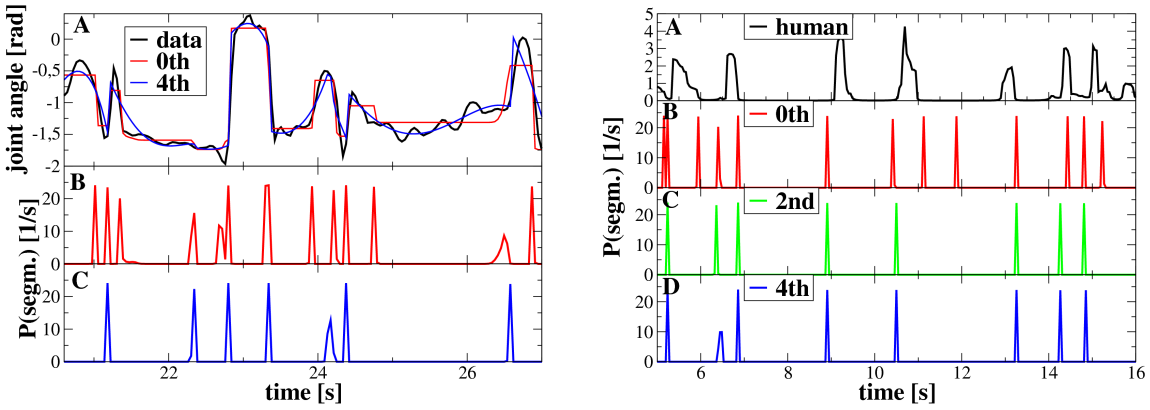

Fig. 3. **Left**: A: fitting a part of a joint angle trajectory with Bayesian binning. Joint angles have not been wrapped around at $-\pi$ to avoid creation of artificial segmentation points. Shown are predictive joint angles with a 0th order (i.e. bin-wise constant) and a 4th order observation model (see section 4). B,C: predictive segmentation densities for these two observation models. The 0th order model needs more segmentation points to fit the data, and the fit is less faithful than the 4th order model. **Right**: comparison of human segmentation densities with Bayesian binning. Shown is an interval with a few, relatively clear segmentation points and good agreement between human subjects. Note that the human segmentation density (panel A) peaks usually closely to a peak in the density obtained by Bayesian binning. The 0th order model (panel B) predicts more segmentation points than the higher-order models (panels C,D), and the higher-order models are in better agreement to the human segmentation, both in number and location of the segmentation points.

and compute the marginal likelihood of the data, i.e. the denominator of eqn.(30), noting that the last two factors of the numerator (eqn.(31)) have the same functional form (with respect to $\mathbf{P}$) as the Wishart prior, but with the prior parameters replaced by the posterior parameters $\hat{\nu}$ and $\hat{\mathbf{V}}$:

$$p(D|\mathbf{A},\mathbf{B},\nu,\mathbf{V}) = \frac{\sqrt{|\mathbf{B}|}^{L}}{\sqrt{|\hat{\mathbf{B}}|}^{L}} \sqrt{2\pi}^{-L} Z_w(\nu,\mathbf{V})^{-1} Z_w(\hat{\nu},\hat{\mathbf{V}}). \tag{34}$$

Thus, the posterior of $\mathbf{P}$ given D is:

$$p(\mathbf{P}|\nu,\mathbf{V},D) = Z_w(\hat{\nu},\hat{\mathbf{V}})^{-1} |\mathbf{P}|^{\frac{\nu-L-1}{2}} \exp\left(-0.5\ \mathbf{tr}\left[\hat{\mathbf{V}}^{-1}\mathbf{P}\right]\right) \tag{35}$$

which is a Wishart density, whence conjugacy is established. Furthermore, eqn.(34) is the marginal likelihood required by BB for each bin.

5. RESULTS

To determine the segmentation densities, we applied BB to joint angle trajectories of combinations of shoulder, elbow and knee angles, since these joints might be expected to be particularly expressive in taekwondo motions. Fig. (3), left, panel A shows the predictive trajectories computed with a 0th order and a 4th order observation model. Both models fit the data well, but the 4th order model yields a better fit with fewer bin boundaries. Panels B and C in fig. (3), left, depict the predicted segmentation densities. The 4th order model not only uses less bin boundaries, it also results in a more clear-cut segmentation.

Fig. (3), right, shows comparisons between human segmentation densities and those obtained by BB. Note that the human segmentation density (panel A, in fig. (3), right) peaks usually close to a peak in the density obtained by Bayesian binning. The 0th order model (panel B) over-segments, this over-segmentation is much reduced for the higher-order models (panels C and D). On closer inspection, one can see that the BB density

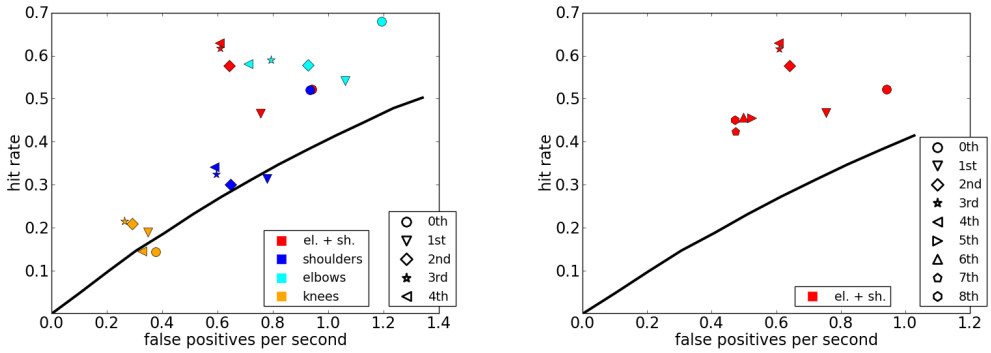

Fig. 4. Hit rate analysis. Joint angle (combinations) are indicated by colour ('el.+sh': elbow and shoulder angles segmented simultaneously). The polynomial order of the observation model (eqn.(9)) corresponds to the marker shape. *Black solid lines* are 'lines of no discrimination' generated by sampling (machine) segmentation points from a homogeneous Poisson process. *Left*: jointly segmenting elbow and shoulder angles with a 3rd or 4th order observation model yields the best compromise between a high hit rate and a low number of false positives per second. While the hit rate of a 0th order model segmenting elbows is slightly higher, it produces ≈ twice the number of false positives per second. *Right*: increasing the polynomial order beyond 4 leads to a significant decrease of the hit rate. For details, see text.

sometimes peaks at the beginning of a (broader) peak of the human density, and sometimes more towards the end. This can be attributed to the periods of stillness between some of the individual taekwondo action 'atoms': humans tend to place a segmentation boundary somewhere in that still period, whereas BB will tend to segment either at the beginning or the end of it. This ambiguity could be resolved by biasing the prior on the polynomial coefficients (eqn.(14)) towards zero velocity at the corresponding lower bin boundary.

For a more quantitative performance evaluation, we conducted a hit rate analysis, see fig. (4). We computed the data in these graphs as follows: assume that both the BB segmentation points (BBSP) and the average human segmentation points (HSP) were known with certainty. Then, a BBSP counts as a hit if it is within a ±250ms accuracy window[2] of a HSP, and if no other BBSP has been assigned to that HSP already. All remaining BBSPs comprise the false positives. HSPs without a matching BBSP count as misses. The hit rate is then computed in the usual way:

$$\text{hit rate} = \frac{\text{hits}}{\text{hits} + \text{misses}}$$

Computing a false positive rate for a standard ROC analysis

$$\text{false positive rate} = \frac{\text{false positives}}{\text{false positives} + \text{true negatives}}$$

is somewhat problematic, since it requires the evaluation of the "true negatives", i.e. the number of instances where there is neither a BBSP nor an HSP. This number depends on the chosen discretisation: the false positive rate can be reduced almost arbitrarily by increasing the temporal resolution, since both BBSPs and HSPs are (almost) point events. We sidestep this issue by evaluating the false positives per second instead, which is largely independent of the temporal resolution. Since we do not know the HSPs and BBSPs with certainty, but only their densities, we compute expected values for the hit rate and the false positives per second by sampling segmentation points from both densities until the standard error of the expectation estimates is $\leq 10^{-3}$. Also, we thresholded the human segmentation density at 2 events/second to remove

[2]The time window was defined as ±250ms because healthy humans between 20 and 30 years show mean reaction times of about 200 ms to visual stimulation

unclear segmentation points (i.e. where the human segmenters were mostly in disagreement). As a reference for the expected hit rates and false positives per second when guessing randomly, we computed a "line of no discrimination" (solid lines in fig. (4)). This line is computed by drawing uninformative segmentation events from a homogeneous Poisson process with rate parameter λ. Each setting of λ corresponds to one point on the line of no discrimination.

In fig. (4), left, we segmented joint angle trajectories from shoulders, elbows, knees and elbows+shoulders (el.+sh.) with observation models having polynomial orders between 0 and 4. In this figure, marker shape indicates polynomial order, while col or denotes joint angle. The priors on the parameters were initialised to mean 0 and a diagonal covariance matrix with variance 160. The segmentation obtained from the knees is largely uninformative, whereas most polynomial orders provide an informative signal about when segmenting arm joints. The best compromise between high hit rate and low number of false positives per second is achieved with orders 3 or 4, segmenting elbows and shoulders together. Increasing the polynomial order beyond 4 does not seem to improve the results, as can be seen in fig. (4), right: while the number of false positives per second keeps decreasing with increasing order > 4, the hit rate drops off as well.

The fact that models with order 3 or 4 provide a better match than the lower orders indicates that humans employ (the visual equivalent of) angular acceleration discontinuities, rather than discontinuities in angular velocities when segmenting action streams. This agrees with the 'minimum jerk' hypothesis [Flash and Hogan 1985].

6. CONCLUSION

We presented two novel contributions in this paper: firstly, we have extended Bayesian binning by employing piecewise polynomial observation models and demonstrated its usefulness for action stream segmentation. Secondly, we have created a benchmark data set for the evaluation of machine segmentation methods compared to human observers.

One limitation of our model is the independence assumption between different bins (see eqn.(1)), implying that human actions can be viewed as a sequence of clearly separable atoms. Our results indicate that this assumption is at least approximately fulfilled for a taekwondo *hyeong*. However, in other type of human action, it might become necessary to model dependencies between data in different bins explicitly. Consider e.g. pair dancing, where the transition from one figure to the next is typically more continuous than in taekwondo. Another example is sign language production, where co-articulation [Segouat and Braffort 2009] induces dependencies between neighbouring signs. BB could be extended to include such dependencies between segments. As long as the factor graph (see fig. (2)) remains singly connected, exact inference will still be possible, albeit with a higher computational effort.

[Polyakov et al. 2009] successfully fitted trajectories with parabolic pieces, we showed that higher orders yield a yet better agreement with human psychophysical data. One could go even further and use a hidden Markov model (HMM) in each bin.u Switching HMMs were used before for action segmentation [Green 2003], a BB prior on top of HMMs might be a feasible way of switching between them.

In [Barbič et al. 2004], three methods for motion capture data segmentation are compared. The methods are based on segment-wise PCA, probabilistic PCA (pPCA) and finite Gaussian mixtures. The approach based on pPCA yielded the best results. Since our observation model (eqn.(6)) learns a full covariance matrix from the data for each segment, a pPCA decomposition of the data in each segment could be extracted from the posterior parameters of the observation model. Conversely, our 0th order observation model could be viewed as a marginalised pPCA model with constant mean per segment, which is the pPCA observation model of [Barbič et al. 2004]. As illustrated in fig. (4), our higher-order models offer a significant performance advantage over a pPCA model with constant means on our data.

Another way of extending our approach would be to include context information into the segmentation process, in addition to the purely kinematic information currently used. [Zacks et al. 2009] reports that

humans use context information for segmentation tasks when such is available, and rely increasingly on kinematics when context is reduced.

Acknowledgements

This work was supported by EU projects FP7-ICT-215866 SEARISE, FP7-249858-TP3 TANGO, FP7-ICT-248311 AMARSi and the DFG. We thank Engelbert Rotalsky, Hans Leberle and the Taekwondo Unions of Nordrhein-Westfalen and Baden-Württemberg for cooperation on the data acquisition. We thank S. Cavdaroglu, W. Ilg and T. Hirscher for their help with data collection and post-processing.

REFERENCES

Agam, Y. and Sekuler, R. 2008. Geometric structure and chunking in reproduction of motion sequences. *Journal of Vision 8,* 1, 1–12.

Arikan, O. and Forsyth, D. A. 2002. Interactive motion generation from examples. *ACM Trans. Graph. 21*, 483–490.

Barbič, J., Safonova, A., Pan, J.-Y., Faloutsos, C., Hodgins, J. K., and Pollard, N. S. 2004. Segmenting motion capture data into distinct behaviors. In *Proceedings of Graphics Interface 2004*. GI '04. Canadian Human-Computer Communications Society, School of Computer Science, University of Waterloo, Waterloo, Ontario, Canada, 185–194.

Bishop, C. M. 2007. *Pattern Recognition and Machine Learning*. Springer.

Dickman, H. R. 1963. The perception of behavioral units. In *The stream of behavior*, R. G. Barker, Ed. Appleton-Century-Crofts, New York, 23–41.

Endres, D. and Földiák, P. 2005. Bayesian bin distribution inference and mutual information. *IEEE Transactions on Information Theory 51,* 11, 3766 – 3779.

Endres, D. and Oram, M. 2010. Feature extraction from spike trains with bayesian binning: latency is where the signal starts. *Journal of Computational Neuroscience 29*, 149–169.

Fearnhead, P. 2006. Exact and efficient bayesian inference for multiple changepoint problems. *Statistics and Computing 16,* 2, 203–213.

Flash, T. and Hogan, N. 1985. The coordination of arm movements: an experimentally confirmed mathematical model. *J. Neurosci.* 5, 1688–1703.

Green, R. D. 2003. Spatial and temporal segmentation of continuous human motion from monocular video images. In *Proceedings of Image and Vision Computing*. New Zealand, 163–169.

Hutter, M. 2007. Exact bayesian regression of piecewise constant functions. *Journal of Bayesian Analysis 2,* 4, 635–664.

Ilg, W., Bakir, G., Mezger, J., and Giese, M. 2004. On the representation, learning and transfer of spatio-temporal movement characteristics. *International Journal of Humanoid Robotics 1,* 4, 613–636.

Kovar, L., Gleicher, M., and Pighin, F. 2002. Motion graphs. *ACM Trans. Graph. 21*, 473–482.

Kschischang, F., Frey, B., and Loeliger, H.-A. 2001. Factor graphs and the sum-product algorithm. *IEEE Transactions on Information Theory 47,* 2, 498–519.

Marida, K. V. and Jupp, P. E. 2000. *Directional Statistics*. Wiley.

Newtson, D. and Engquist, G. 1976. The perceptual organization of ongoing behavior. *Journal of Experimental Social Psychology 12,* 5, 436 – 450.

Petersen, K. B. and Pedersen, M. S. Nov. 14, 2008. The matrix cookbook.

Polyakov, F., Stark, E., Drori, R., Abeles, M., and Flash, T. 2009. Parabolic movement primitives and cortical states: merging optimality with geometric invariance. *Biol. Cybern. 100,* 2, 159–184.

Roether, C. L., Omlor, L., Christensen, A., and Giese, M. A. 2009. Critical features for the perception of emotion from gait. *Journal of Vision 9,* 6, 1–32.

Rose, C., Bodenheimer, B., and Cohen, M. F. 1998. Verbs and adverbs: Multidimensional motion interpolation using radial basis functions. *IEEE Computer Graphics and Applications 18*, 32–40.

Safonova, A., Hodgins, J. K., and Pollard, N. S. 2004. Synthesizing physically realistic human motion in low-dimensional, behavior-specific spaces. *ACM Trans. Graph. 23*, 514–521.

Segouat, J. and Braffort, A. 2009. Toward the study of sign language coarticulation: Methodology proposal. In *Proceedings of the 2009 Second International Conferences on Advances in Computer-Human Interactions*. ACHI '09. IEEE Computer Society, Washington, DC, USA, 369–374.

Shipley, T. F., Maguire, M. J., and Brumberg, J. 2004. Segmentation of event paths. *Journal of Vision 4,* 8.

Zacks, J. M., Kumar, S., Abrams, R. A., and Mehta, R. 2009. Using movement and intentions to understand human activity. *Cognition 112,* 2, 201 – 216.