Human TECHNOLOGY

ISSN: 1795-6889

https://ht.csr-pub.eu Volume 18(3), December 2022, 250-266

SEGMENTATION BOUNDARIES IN ACCELEROMETER DATA OF ARM MOTION INDUCED BY MUSIC: ONLINE COMPUTATION AND PERCEPTUAL ASSESSMENT

Juan Ignacio Mendoza G.
University of Jyväskylä
Department of Music, Art and Culture Studies
Finland

Abstract: *Segmentation is a cognitive process involved in the understanding of information perceived through the senses. Likewise, the automatic segmentation of data captured by sensors may be used for the identification of patterns. This study is concerned with the segmentation of dancing motion captured by accelerometry and its possible applications, such as pattern learning and recognition, or gestural control of devices. To that effect, an automatic segmentation system was formulated and tested. Two participants were asked to 'dance with one arm' while their motion was measured by an accelerometer. The performances were recorded on video, and manually segmented by six annotators later. The annotations were used to optimize the automatic segmentation system, maximizing a novel similarity score between computed and annotated segmentations. The computed segmentations with highest similarity to each annotation were then manually assessed by the annotators, resulting in Precision between 0.71 and 0.89, and Recall between 0.82 to 1.*

Keywords: *gestural interface, perceptual evaluation, temporal segmentation, accelerometer, bodily motion, similarity*

DOI: https://doi.org/10.14254/1795-6889.2022.18-3.4

INTRODUCTION

The advancement in miniaturization of accelerometers, gyroscopes and magnetometers has made it possible to develop portable and wearable systems that sense the movement of the human body. This has opened doors for many applications in a vast range of domains. Many such applications require identifying segmentation boundaries within movement, that is, where data changes from one regime to another. Following this, the detected segments can be classified or clustered. Some methods detect segmentation boundaries in the same process that performs classification or clustering. Examples of applications that use these processes include systems for detecting, recognizing and monitoring activities for clinical diagnosis or assisting in sports training (Cornacchia, Ozcan, Zheng, &Velipasalar, 2017).

The focus of the current study was to identify segmentation boundaries within the movement of a person dancing. In a practical application, the detected segmentation boundaries may be used to control playback of sound, music or lighting, for example. The movement of the dancer may be sensed in a number of different ways, but this study focuses on the use of a single triaxial accelerometer. The output is the time when a segmentation boundary has occurred, with respect to real time. Then, this information may be used for the control of a separate process (e.g., triggering events) or for machine-learning processes such as clustering or classification of the found segments.

It is desirable that the result of the segmentation system is produced fast enough for near-real-time interaction. Also, it is necessary that the motion segments are meaningful to an observer. In other words, motion segments produced by the system should match the segments perceived by an observer. The meaningfulness of motion segments would additionally facilitate the learning of motion patterns and mappings to audio or visual effects. To that extent, it must be acknowledged firstly, that human perception of bodily movement is highly subjective (Bläsing, 2015; Kahol, Tripathi, & Panchanathan, 2004; Zacks, Kumar, Abrams, & Mehta, 2009) and is hierarchically structured such that short patterns are grouped into larger ones (Bernard, Dobermann, Vögele, Krüger, Kohlhammer, & Fellner, 2017; Dreher, Kulp, Mandery, Wächter, & Asfour, 2017; Krüger, Kragic, Ude, & Geib, 2007; Lin, Karg, & Kulić, 2016). Also, it must be taken into consideration that dance patterns may or may be not repetitive. Thus, the system must be capable of detecting repetitive and non-repetitive patterns, and must allow the user to make adjustments to obtain perceptually meaningful results.

The algorithm described by Foote (2000) for segmentation of digital audio was found to be an appropriate candidate for segmentation of dance movement. This algorithm has subsequently been used for segmentation of video (Foote & Cooper, 2003), and of dance motion based on speed extracted from video (Tardieu et al., 2009). It has also been used to identify boundaries between activities such as walking, jogging and sitting, in single-axis accelerometer data (Rodrigues, Probst, & Gamboa, 2021). While most published implementations are online (i.e., data is processed serially as it is input to the algorithm), Schätti (2007) described an online implementation for segmentation of an audio signal. Also these implementations have been tested on data whose segments span several seconds or minutes (e.g., sections of a song, walking). Therefore, the current study has focused on the adaptation of an online version of the algorithm to work with a triaxial accelerometer signal, and the assessment of its capability to meet the requirements of the intended application. The

contributions of the present study are, first, the application and testing of the segmentation algorithm at a smaller time-scale (i.e., short dancing patterns spanning a few seconds), and a more robust perceptual assessment than those used in previous work. The second contribution is a novel measure to evaluate the similarity between computed and perceived segmentation boundaries.

This report is structured as follows: The remainder of the introduction presents a succinct review of the state-of-the-art methods that most closely meet the requirements stated above, including unsupervised near-real-time detection of segmentation boundaries, boundaries of self-similarity checkerboard patterns, and assessments of effectiveness. In favor of a timely report, a comprehensive comparison of different techniques is out of the scope of this study. Following this, the Methods utilized and the Results so obtained are reported. Finally, the Conclusion provides a summary of the study, including directions for future work.

Unsupervised Near-Real-Time Detection of Segmentation Boundaries

Several algorithms that detect segmentation boundaries and give results in near-real-time have been tested with data from accelerometers. For example, Gharghabi et al. (2019) described a method that evaluates the similarity in shape –but not in statistical properties– between all fixed-length windows within a bigger window, the length of which is specified by the user. A segmentation boundary is recorded where the similarity is minimal. This method assumes that each segment will be composed of at least two instances of a periodic motion.

Another approach is to pose the task as a multivariate change-point detection problem (Endres, Christensen, Omlor, & Giese, 2011; Gong, Medioni, & Zhao, 2014; Krüger et al., 2017; Zhou, De la Torre, & Hodgins, 2012). Essentially, a change-point indicates a difference in statistical properties of the data within a sliding window (Aminikhanghahi & Cook, 2017; Fathy, Barnaghi, & Tafazolli, 2018; Liu, Yamada, Collier, & Sugiyama, 2013; Patterson et al., 2016). The sliding window is a free parameter that adjusts time-scale (i.e., granularity). Depending on the method, other free parameters may need to be adjusted. Zameni et al. (2020) described a method that efficiently identifies segmentation boundaries in signals that can be highly dimensional. This method has initialization parameters, but no parameters that can be used to explicitly adjust time-scale or relevance. The cited systems were tested with various types of data. When the test data had been recorded by triaxial accelerometers, the tests aimed to segment activities that take at least a few seconds to complete. However, segments of dancing motion may range from less than a second to more than a few seconds.

Boundaries of Self-Similarity Checkerboard Patterns

The detection of change-points in motion data can be seen as equivalent to novelty detection, which is the identification of abrupt changes in data by a system, without training of the system (Markou & Singh, 2003). Foote (2000) described a method suitable for finding segmentation boundaries in musical audio signals. This method exploits the characteristic checkerboard patterns that can be observed in a self-similarity distance matrix of audio features through time, by correlating a checkerboard kernel along the diagonal of the matrix. This results in a novelty score that indicates the rate of change in the data. The peaks of the

novelty score indicate change-points that correspond to perceived changes in the music. The granularity of the novelty score is adjusted with the width of the kernel and relevant peaks can be selected over a threshold.

Assessment of Effectiveness

To measure the effectiveness of segmentation algorithms, most published studies have relied at least to some extent on classic measures of precision, recall and accuracy, by comparing human-annotated ground truth boundaries annotated by one or more people with computed boundaries. These measures work well for classification problems in which the options are either "match" or "not a match" between a computed boundary and a ground truth boundary. Dreher et al. note that a computed segmentation boundary being only slightly different to the ground truth should be counted as a match. This is usually solved by establishing a window around each ground truth boundary. A computed point is deemed to be a true positive if it lies within that window. This approach was used in the study by Zameni et al., for example. Dreher et al. proposed a method that involves a window weighted with a normal distribution. However, the problem with this approach is that the window's width is fixed while there is no certainty that any given width will correspond to the true probability distribution for the occurrence of a boundary, for all boundaries. It is not possible to generalize the temporal length of the transition from one motion to another. In contrast, the evaluation method used by Gharghabi et al. consists of a score that measures the temporal distance between each computed boundary and the closest boundary in the ground truth. All the distances are added and then divided by the total time. However, this score does not penalize extra or missing computed boundaries, which is problematic as there is no certainty that the number of annotated and computed boundaries will always be the same. Lin et al. (2013) describe another approach for evaluation of results, in which all frames in the ground truth segments are labelled and the number of frames in the computed segments corresponding to the ground truth-labels constitute the measure of similarity. This last method might be appropriate for classification of segments but it might be too restrictive for evaluating only the boundaries. This is because boundaries of short false-positive computed segments (e.g., transitions between motions) will break the continuity of parallel labelling resulting in a very high dissimilarity score. Mendoza (2014), and also Mendoza and Thompson (2017), proposed similarity scores that measure the distance between ground truth and computed boundaries as in the method by Gharghabi et al., but also penalize missing or extra computed boundaries.

The Present Study

The following section describes the implementation of Foote's algorithm for the segmentation of accelerometer data. Then, an experimental assessment is described in which ground truth is used to tune the algorithm's free parameters using a revised version of the similarity measure by Mendoza and Thompson. In contrast to previous studies, the computed results are not assessed by means of a similarity measure but manually by the same annotators who provided the ground truth.

METHODS

Detection of Segmentation Boundaries

This subsection describes the method for finding temporal segmentation boundaries, focusing on its online implementation and its adaptations to work with accelerometer data. A succinct description of the original offline version is provided. For details of the algorithm in general and the offline version, the reader is directed to the original source (Foote, 2000).

The offline version of the algorithm has as input data stored in memory, which has been sampled at regular intervals. This data is represented by the matrix $M \in \mathbb{R}$, so that $M_{1:m} = [F_1, F_2 \dots F_m]^{\mathrm{T}}$. Each frame F at time-index $t \in \{1 \dots m\}$ contains data for each sample. A distance matrix $D \in \mathbb{R}^{m \times m}$ is computed for all data in M. D is a self-similarity matrix. A two-dimension checkerboard kernel is produced by the Kronecker product of checkerboard matrix C and only-ones matrix J of width n as follows:

$$C = \begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix} \tag{1}$$

$$K = C \otimes J \tag{2}$$

K is then tapered by multiplying it element-wise with a two-dimensional Gaussian (i.e., a normal distribution). Next, K is correlated along the diagonal of D. The result of this correlation is novelty score N, the peaks of which indicate the locations of segmentation boundaries. The peaks can be selected by a threshold θ, discarding peaks of lower values that might be irrelevant. Hence, n and θ are free parameters for granularity and peak relevance, respectively.

The online version of the algorithm consists in M being a stream of data frames $F_t = (f_x^t, f_y^t, f_z^t)$, sampled at regular intervals, containing the three axes of the accelerometer. A window of n frames is stored in a buffer W_{nov} (Figure 1a). For each incoming frame, the last frame in the buffer is removed while the current frame is stacked in the first position, and distance matrix $D \in \mathbb{R}^{n \times n}$ is computed for W_{nov} (Figure 1b). In this study, Euclidean distance was used. Then, the inner product between Gaussian-tapered checkerboard kernel K and D is computed, resulting in a new point in novelty score N (Figure 1c).

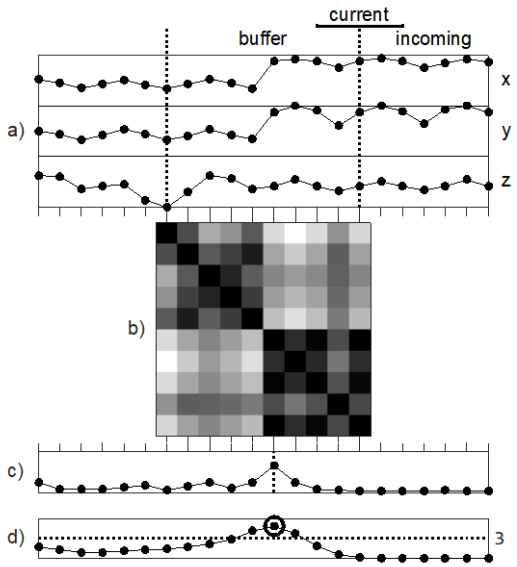

Figure 1. Online detection of temporal segmentation boundaries. Horizontal axes represent time. (a) is triaxial accelerometer data. (b) is self-similarity matrix D of data in the buffer W_{nov}, where lighter shades represent more distance. (c) is novelty score N, where the vertical dotted line indicates the current result. (d) is the smoothed novelty score N', where θ is a threshold and the point in a circle is the selected peak indicating a segmentation boundary. Note that this visualization shows N and N' aligned in time, but in practice there will be a lag due to the low-pass Gaussian filter and the test for a peak.

When tested, Ncontained many irrelevant peaks. Therefore a low-pass filter was applied. The filter used in this study was a one-dimension Gaussian kernel with minima zero and unit area to prevent artefacts at borders and to preserve scale, respectively. This filter is computed upon a second buffer W_{filt} having the size of the one-dimensional Gaussian n_{filt}, resulting in a smooth novelty score N'. Finally, if the current novelty score value is a peak over threshold θ, it is considered a segmentation boundary (Figure 1d). Identification of peaks requires another buffer of only three samples to test a local maximum. Hence, the identification of a novelty peak has lag

$$l = \frac{n+ n_{filt}}{2} + 3 \tag{3}$$

with respect to the current incoming frame.

Since self-similarity matrix D is symmetric, it is necessary to compute only half of it, either the upper or lower triangle, without the diagonal. Also there is no need to compute the whole triangle for each new frame. It is only needed to initialize matrix D with allocation values (e.g., zeros), then compute the distance between the current frame and all the other frames in the buffer. Then, compute the inner product of the upper or lower triangle of D and the corresponding triangle of K. This will output the current novelty value. Then the values within D are shifted, discarding the distances between the oldest frame and the newer ones.

This operation reallocates memory indexes, which takes much less computation time than redundant computation of distance.

The time-scale of the segments may be adjusted dynamically with parameters n and θ. This may be accomplished by fixing the ratio between parameters n_{nov} and n_{filt}, so that parameter n modifies the size of buffers W_{nov} and W_{filt} at the same time. When changing n, a new checkerboard kernel may be computed, or a kernel may be selected from many that might have been previously computed and stored in memory. Because of the operations on D and K, the asymptotical memory complexity is $O(n^2)$ while computing-time complexity is linear. However, in practice n may not grow too much to present a concern, as its size would be limited to the intended granularity and may be reduced by reducing the sampling rate.

Accelerometer Data Collection

Two participants, one female and one male, provided motion data to test the segmentation method. This data was collected at the motion-capture laboratory of the department of Music, Art and Culture Studies at the University of Jyväskylä. These participants are referred to as *dancers* to differentiate them from the participants that provided data for the ground truth and perceptual assessment (see subsection "Ground truth annotation").

In individual sessions, the dancers were asked to "dance with one arm" while holding with the corresponding hand a Nintendo Wii-remote controller. They were asked to move to the music, without displacement of the body, and always facing one corner of the room. While these conditions may not generalize to all dancing scenarios, they provided a clear view of the moving arm to a video camera. Video recordings were later used for manual annotation. The elimination of the random variable of orientation facilitated the annotation task. Also it simplified the analysis, thus making it possible to focus on first solving the segmentation problem in a simple condition before embarking on a more complex scenario. The dancers were told that other than these constraints, they could move as they wanted.

Three musical stimuli were presented through loudspeakers:

1. "Minuet" (Petzold, ca. 1725) MIDI rendition with piano sound, from beginning to end (104 bars, duration 92.5 s.) with no fade-in or fade-out. It has a ternary metre (3/4, or three beats per bar). Both participants declared to know this piece.
2. "Ciguri" (Otondo, 2008) from 56 to 183.7 s. (duration 122.7 s.) with fade-out the last 5 s. This is an electroacoustic piece that has no perceivable beat and therefore no metre. Both participants declared to not know this piece.
3. "Stayin' Alive" (Gibb, Gibb, & Gibb, 1977) from the beginning to 108.5 s. with fade-out the last 2.3 s. It has a binary metre (4/4, or four beats per bar). Both participants declared to know this piece.

The number of performances amounted to six. This was deemed enough for this study as they provided variety: musical genre, metre, familiarity and the gender of the participants. These characteristics would permit to observe to some extent their effect on the test. Furthermore, later these performances were used for the task described in the next section

("Ground truth annotation"). More performances would have extended the annotation task implying the risk of abandonment or fatigue, the latter reducing the reliability of results.

Stimuli were presented in the order listed above and each stimulus was presented twice. During the first presentation, participants were asked to move freely within an area of about $4m^2$, to familiarize themselves with the stimulus. For the second presentation, participants were asked to dance with one arm as described above. Data of the performances were recorded as follows:

- *Accelerometer:* The Nintendo Wii-remote has a triaxial accelerometer, which transmits data in real-time via Bluetooth. This stream was received and recorded by a computer at a rate of 100 Hz, using custom-made software.
- *Video:* A digital video camera recorded video showing the participant's whole body against a white wall. Both participants used their right arm, and were recorded so the image clearly showed the moving arm.
- *Audio:* Digital audio was captured by the microphone of the video camera and by a microphone hanging from the ceiling. The latter was recorded to a digital audio workstation synchronized with the recording of accelerometer data. These signals were subsequently used to synchronize video and accelerometer data.

Ground Truth Annotation

Six participants (3 male, 3 female) were recruited to identify segmentation boundaries in the one-arm-dancing videos. None of them had participated in the data collection described in the previous section. Their ages ranged from 26 to 34 years, with a median age of 27. All were non-Finnish international students at the University of Jyväskylä. All had completed at least an introductory course in music psychology, covering an introduction to perception and segmentation. These participants are referred to as *annotators*, to differentiate them from the *dancers* who performed the one-arm dance (see subsection "Accelerometer data collection").

Each annotator, in an individual session, was asked to watch the videos and identify segmentation boundaries in two conditions. In the first condition, the videos with audio were presented by a computer running custom-made software. The annotators were instructed to press a key when a boundary was identified, in real time. The time of the key relative to the video was recorded by the computer. They had only one chance to perform the task. It was thought that the music in the video may influence the responses as auditory cues, such as pitch or rhythm, and could be used to judge the existence of a boundary. For the second condition, the videos without audio were presented by the computer running a digital audio editor software. In this condition, participants could freely play the video, pause, scroll forward and backwards, place markers and adjust the location of the markers until they were satisfied. In this condition, the annotators did not have a limit of time for the task and the annotation was based solely on visual information.

The following were the instructions to the annotators, common for both conditions:

"You will be presented with six videos, each lasting around two minutes. Each video shows a person 'dancing' with an arm. When doing this, the person does distinct patterns with the arm. A pattern is composed by one distinct movement or several repetitions of the

same movement. When the video is playing press the space bar to indicate a change in pattern. Focus in the movement of the arm holding the white device (it is a sensor).”

The two annotation conditions represented different approaches for perceived segmentation. To assess their suitability, the annotators were interviewed after completing the tasks. They were asked to verbally express what they considered to be difficult or easy about the tasks. All participants mentioned that, in the real-time annotation task, their responses might have been influenced by the music and they were less precise than in the non-real-time condition. The reasons mentioned for this included that in the real-time condition the responses might have been anticipated as an effect of the music. Also, it was mentioned that, in the real-time task, it was more difficult to press the button exactly at the intended time, thus preventing a response to be recorded accurately or in some cases at all. All participants expressed that the non-real time condition allowed for more precise responses, as they could take time to revise them. Because of this, the data relating to real-time audiovisual annotation was deemed inappropriate for use as a ground truth. Thus, non-real-time visual annotation was chosen as ground truth for perceived segmentation boundaries.

Optimization using similarity based on distance and rate of paired elements

A grid search was performed to maximize the similarity between annotated (ground truth) and computed segmentation boundaries, by modification of parameters *n* and θ. This search was performed independently for each accelerometer recording and their corresponding annotations, mimicking the adjustment that might be achieved manually by an end-user or automatically by a machine-learning procedure. Similarity was evaluated by distance and penalization of extra or missing boundaries, improving previous work (Mendoza, 2014; Mendoza & Thompson, 2017).

Consider vectors a and b containing the time indexes of annotated and computed segmentation boundaries, respectively. L is the length, in samples, of the corresponding recorded data, from the start to the end of the musical stimulus. n_a and n_b are the number of boundaries, or length, of a and b respectively. In any case $n_a \geq n_b$ or vice-versa. Each element in a is paired to the closest element in b, so that a' and b' are vectors containing only the paired elements and have equal lengths n_p (equivalent to the shortest between n_a and n_b). Then, the following measures are computed:

Closeness:

$$c = 1 - \frac{1}{L}\sum_{i=1}^{n_p}|a'_i - b'_i| \quad (4)$$

Rate of paired elements:

$$p = \frac{2n_p}{n_a + n_b} \quad (5)$$

Similarity:

$$S = c \cdot p\ , 0 \geq S \geq 1 \quad (6)$$

The distance between paired boundaries is the absolute time difference, as shown in equation 4. Note that two boundaries of either sequence (a or b) may be paired with a single boundary in the other sequence if their distances are equal. Also, if n_a and n_b are not equal and there are no equidistant boundaries to compensate for that inequality, then some boundaries will not be paired and this will be penalized by the rate of paired elements (equation 5). A Monte Carlo simulation was computed with pseudo-random a and b, for $L = 1000$, with n_a and n_b in the range $\{1 \ldots L-1\}$, and 10^4 iterations. The distribution for the resulting S values has an upper p-value of 0.05 at $S = 0.66$.

Perceptual Assessment

The perceptual assessment was made by the same annotators that provided the ground truths. For each annotator, the annotated and computed boundaries with highest similarity were selected. This means that the assessment is for the 'best case scenario'. For each of these sequences of boundaries a video was produced embedding a scrolling timeline with consecutive numbers for boundaries into the corresponding video that was annotated (Figure 2).

Three videos were produced for each annotator. One had markers for their original annotation, to measure the extent of agreement they would have with the annotation they had previously made. A second video had markers for the computed boundaries. A third video had a confounding sequence of boundaries produced by placing a marker in the middle of the segments bounded by the average point for each pair of paired annotated and computed boundaries. The videos with confounding boundaries were intended to reduce the chance of annotators realizing that one of the sequences was their own annotation, and the responses to those videos were not analyzed.

Figure 2. Example frame of a video shown to an annotator for perceptual assessment. The same video without the numbered markers had been used for annotation.

The videos contained no audio, as the annotations used in the computation of boundaries corresponded to video without audio. Each video was embedded in a webpage and had on-screen controls that could be activated with a pointing device (e.g., mouse, trackpad) to play, stop, scroll forward and backwards. The pages were presented in random order by an automatic system that also recorded responses. Each page consisted of instructions, the video and a list of numbered items, one for each marker. Each item in the list had two buttons that could be selected by clicking on them. One button was to answer "yes, there is a change in pattern" and was recorded as a *confirmed boundary*. The other button was to answer "no, there is no change in pattern" and was recorded as a *rejected boundary*. This assessment is used in replacement of the paradigm used in previous studies that considered a computed boundary to be correct if it is within a window around a ground truth boundary. It has the advantage of not needing to specify a fixed window.

The definition of the task was identical to the one given for the annotation task. One distinct questionnaire was produced for each annotator with the corresponding videos. This questionnaire did not reveal how the segmentation sequences were produced. After completing each page all responses were recorded and options were shown to immediately continue to the next page or to continue later. The annotators were asked to complete the questionnaire in their own space and time, using their own computers and to take as much time as they needed.

The decision to assess the best-case-scenario boundaries was made after testing the questionnaire. This test was done with different participants who would take up to 50 minutes to complete a questionnaire with three videos. It was decided that the questionnaire should not exceed three videos, to prevent fatigue and abandonment.

The data obtained from the questionnaires was processed to obtain the following relevance measures:

$$Precision(computed) = \frac{n_{cb}}{n_b} \tag{7}$$

$$Recall(computed) = \frac{n_{cb}}{n_{cb}+n_{ca}-n_p} \tag{8}$$

$$Precision(annotated) = \frac{n_{ca}}{n_a} \tag{9}$$

where n_{cb} is the number of confirmed computed boundaries (true positives), n_b is the number of computed boundaries (true and false positives), n_{ca} is the number of confirmed annotated boundaries, n_p is paired annotated and computed ($n_{ca} - n_p$ is false negatives), and n_a is the number of annotated boundaries (true and false positives). $Precision(annotated)$ may be considered as an indication of the assessment's reliability. It is not possible to obtain $Recall(annotated)$ as false negatives would require the possibility of adding new boundaries, which was not part of the assessment task.

RESULTS AND DISCUSSION

Computation of the grid search was performed with the recorded accelerometer data downsampled to 25 Hz. The standard deviation σ for the two-dimensional Gaussian that tapers *K* and the one-dimensional Gaussian smoothing filter for N' were set to $\sigma = n/5$. The length of the one-dimensional Gaussian was set to *n*; that is, to the width of K and D. The standard deviation of both Gaussians was searched within $\sigma = \{0.5,0.6,\ldots,2\}$ seconds. Since recorded accelerometer data was used, computation was performed in non-real-time. Therefore, the filtered novelty score was rescaled to $0 \geq N \geq 1$ and the threshold for peak selection was searched within $\theta = \{0,0.1,\ldots,0.5\}$. For real-time computation, these values would yield a lag time of l = {0.22, 0.24, ..., 0.52} seconds. Note that lag time does not consider computation time, which depends on the specific computing device used.

The highest lag time among the results is 0.5s, for the segmentation corresponding to Annotator 2, of Dancer 1, to "Minuet". The median lag time was 0.35s. Considering this time scale, this system is not suitable for any practical application that requires immediate perceptual real-time response (i.e., up to about 10 to 50 milliseconds). However, this lag time is suitable for applications in which the occurrence of a segmentation boundary is not to be acted upon immediately. For example, this delayed response may be mapped to a procedure that changes the stimulus music in such a way that it prompts the dancer to change the motion pattern, thus creating a feedback loop. Another use of this delayed response is to record the segments' times, then compute statistics (e.g., mean, standard deviation) and use those for a larger time-scale control of music, lights or other actionable medium. Furthermore, the segmentation result may be used to produce a near-real-time visual or sonic display that may be useful in clinical applications and research in biomechanics, for example.

Tables 1 and 2, respectively, show values for maximum distance d and similarity (S) obtained in the grid search, where $d = |a' - b'|$. The distance is expressed in seconds. The minimum similarity value ($S = 0.56$) has a p-value of 0.39, while the minimum mean similarity value ($S = 0.62$) has a p-value of 0.17. These minimum values represent the worst performance of the automatic segmentation. The greatest mean S values were found for the musical stimuli "Minuet" and "Stayin' Alive", which both have a clear beat and were familiar to the dancers. Conversely, similarity is lower for "Ciguri", which is a piece that has no clear beat and was not familiar to the dancers. This suggests that the effectiveness of the method may be directly related to both or either of these conditions: the presence of a clear beat, and the familiarity the dancers might have with the musical stimulus. Also the table shows that most maxima d seem too large to indicate corresponding paired boundaries. Although this may be considered a limitation of the method, it is still possible that the highly distant computed boundaries are confirmed in the perceptual assessment.

Table 1. Maximum Distance (d) in seconds, between Annotated and Computed Boundaries.

	Dancer 1			Dancer 2		
Annotator	Minuet	Ciguri	Stayin' Alive	Minuet	Ciguri	Stayin' Alive
1	3.80	2.52	2.62	2.33	6.07	2.11
2	4.11	3.59	1.69	5.67	6.74	2.22
3	4.53	7.87	1.96	3.60	5.31	2.84
4	4.44	3.10	3.74	1.15	5.78	3.29
5	3.82	6.31	2.79	2.13	2.27	0.73
6	1.59	2.86	2.72	2.47	1.90	1.56
mean	3.71	4.38	2.52	2.89	4.68	2.13

Table 2. Similarity (S) between Annotated and Computed Boundaries.

	Dancer 1			Dancer 2		
Annotator	Minuet	Ciguri	Stayin' Alive	Minuet	Ciguri	Stayin' Alive
1	0.64	0.66*	0.74*	0.71*	0.75*	0.83*
2	0.76*	0.63	0.68*	0.82*	0.63	0.80*
3	0.71*	0.60	0.68*	0.71*	0.68*	0.91*
4	0.61	0.57	0.74*	0.82*	0.67*	0.74*
5	0.66*	0.68*	0.73*	0.64	0.63	0.71*
6	0.56	0.60	0.70*	0.64	0.61	0.74*
mean	0.66*	0.62	0.71*	0.72*	0.66*	0.79*

* $p \leq 0.05$ (not adjusted for multiple comparisons)

Table 3 contains relevance values for the case of maximum similarity for each annotator. The corresponding sequences of annotated and computed boundaries are visualized in Figure 3. The fifth and sixth boundaries of Annotation 2 seem to be too far for any of them to correspond to the fifth computed boundary. However, this boundary was confirmed in the perceptual assessment. It is not possible to conclude whether this boundary corresponds to any of the annotated boundaries, or if it is a new boundary that was unseen at the annotation task (i.e., serendipity effect) or if it was a mistake made by the annotator in the assessment task.

Table 3. Perceptual Assessment of Annotated and Computed Segmentation with Highest Similarity (S) for each Annotator.

Annotator	**Stimulus**	**Dancer**	$\boldsymbol{S}$	$\boldsymbol{Precision\ (computed)}$	$\boldsymbol{Recall\ (computed)}$	$\boldsymbol{Precision\ (annotated)}$
1	Stayin' Alive	2	0.83	0.75	0.82	1
2	Minuet	2	0.82	0.86	0.86	0.75
3	Stayin' Alive	2	0.91	0.89	1	0.67
4	Minuet	2	0.82	0.71	1	0.71
5	Stayin' Alive	1	0.73	0.71	0.88	0.95
6	Stayin' Alive	2	0.74	0.80	0.92	0.86
mean			0.81	0.79	0.91	0.82

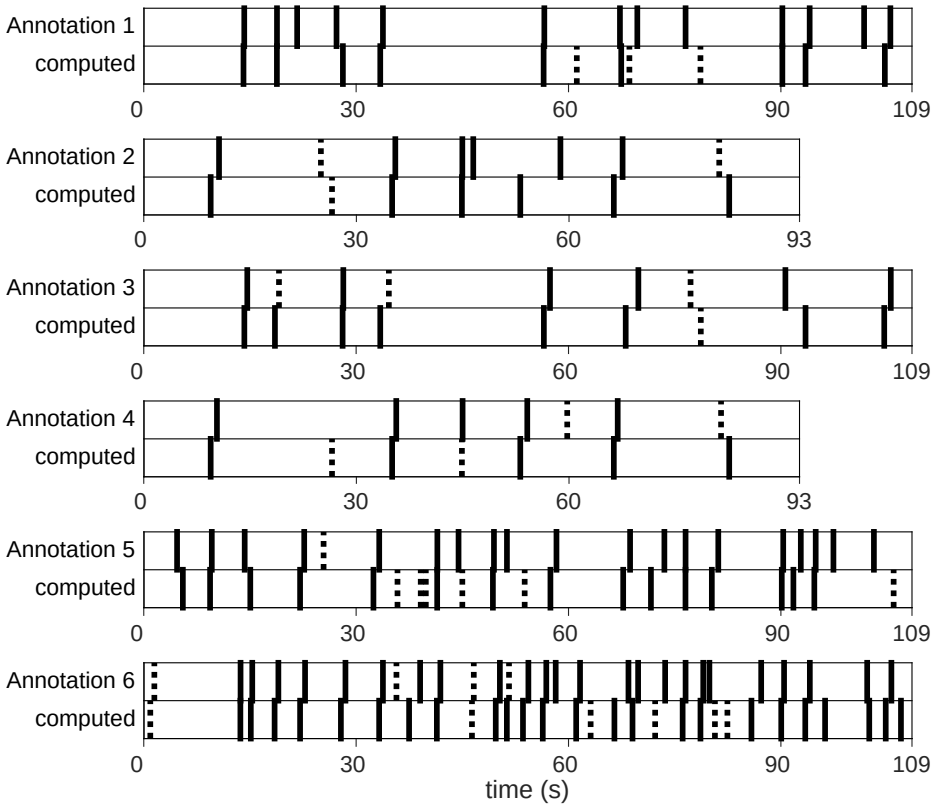

Figure 3. Annotated and closest computed segmentation boundaries for each annotator, corresponding to Table 3. Full lines indicate confirmed and dotted lines indicate rejected.

Another problem is that most annotators rejected boundaries that they had previously annotated, as shown by measure $Precision(annotated)$. While these values are fairly high, some assessment responses look counter-intuitive. For instance, the third boundary of Annotation 4 is evidently close enough to its computed counterpart to be considered an exact match. However, the computed boundary was rejected as shown by the dotted line. Another example that may cast doubt on the perceptual task is the second and fourth boundaries of Annotation 3. These were rejected but their computed counterparts, even being noticeably very near, were confirmed. These odd assessment responses are not the norm, but they raise questions about the reliability of the perceptual tasks.

The two aforementioned assessment problems may be solved by a revised questionnaire including a task that shows both annotated and computed boundaries in the same time line, thus making evident to the annotator the distance between them. In addition, the task would require the annotator to explicitly indicate the corresponding annotated boundary for each computed boundary and vice-versa, if such correspondence exists. Despite the drawbacks of the segmentation and assessment methods, the best-case scenario reveals very high Precision and Recall values. This is relevant as the best-case scenario is akin to the best possible re-tuning that a user could make in a practical application scenario.

A further limitation of this study is that the annotation and assessment tasks were done at different times. This explain the odd responses mentioned above. A possible solution would be to integrate annotation, automatic segmentation, optimization, and assessment, into one procedure.

CONCLUSIONS

This article has presented an adaptation, testing and perceptual assessment of a method to compute segmentation boundaries in accelerometer data. The method is based on an algorithm widely used for segmentation of digital audio (Foote, 2000). Experimental testing of the adapted and extended algorithm used accelerometer data of subjects moving their arm to music, as a simplistic form of dance, from which segmentation boundaries were computed. The fine tuning of the algorithm's parameters was based on annotators' responses, using a novel measure of distance of paired elements between computed and annotated boundaries, combined with penalization for missing or extra boundaries. Perceptual assessment, consisting of rejection or confirmation of computed boundaries, resulted in fairly high values for measures of relevance *Precision* and *Recall*. The segmentation procedure requires a context-dependent minimum time to produce a response, which in this study was maximum about half a second. This is suitable for systems that do not require an immediate response.

Future work on the perceptual assessment of segmentation boundaries should include a task to pair computed and annotated boundaries, in combination with the task to reject or confirm boundaries. It would also be useful to evaluate more and different input data modalities for computing segmentation, as well as manually or automatically learned features that might improve effectiveness. Furthermore, after the segmentation and assessment methods presented in this article are improved as mentioned, they should be incrementally tested on more complex motion and more realistic conditions. Possible next steps might be to attempt segmentation of dancing motion using both arms, legs, the full body, allow free displacement, different musical stimuli and so forth.

IMPLICATIONS FOR RESEARCH AND APPLICATION

This study has developed and tested a system to produce near-real-time segmentation sequences of accelerometer data. This system may be useful for proposing segmentation to a final user, making the process faster than manually. For example, the system could produce several sequences at different granularity levels, out of which the user selects the most appropriate. Likewise, a matrix of multigranular segmentation sequences may be used without any further screening by the user. As such, the system may see a number of practical applications, for example the inspection of data (e.g., identification of daily activity events in data recorded by a wearable accelerometer) or mapping the segmentation results to actionable processes (e.g., gestural control of music, lights, etc.). An important contribution of this study is the formulation of a novel non-parametric similarity measure based on distance and rate of paired elements. Although the measure was developed to assess similarity of segmentation sequences, it may be used to assess the similarity between any pair of sequences of ordered numbers.

REFERENCES

Aminikhanghahi, S., & Cook, D. J. (2017). A survey of methods for time series change point detection. *Knowledge and information systems, 51*(2), 339-367. https://doi.org/10.1007/s10115-016-0987-z

Bernard, J., Dobermann, E., Vögele, A., Krüger, B., Kohlhammer, J., & Fellner, D. (2017). Visual-interactive semi-supervised labeling of human motion capture data. *Electronic Imaging, 2017*(1), 34-45. https://doi.org/10.2352/ISSN.2470-1173.2017.1.VDA-387

Bläsing, B.E. (2015). Segmentation of dance movement: effects of expertise, visual familiarity, motor experience and music. *Frontiers in psychology 5*, 1500. https://doi.org/10.3389/fpsyg.2014.01500

Cornacchia, M., Ozcan, K., Zheng, Y., & Velipasalar, S. (2017). A survey on activity detection and classi cation using wearable sensors. *IEEE Sensors Journal 17*(2), 386–403. http://doi.org/10.1109/JSEN.2016.2628346

Dreher, C. R., Kulp, N., Mandery, C., Wächter, M., & Asfour, T. (2017). A framework for evaluating motion segmentation algorithms. In *2017 IEEE-RAS 17th International Conference on Humanoid Robotics (Humanoids)* (pp. 83-90). IEEE. https://doi.org/10.1109/HUMANOIDS.2017.8239541

Endres, D., Christensen, A., Omlor, L., & Giese, M.A. (2011). Emulating human observers with bayesian binning: Segmentation of action streams. *ACM Transactions on Applied Perception (TAP), 8*(3), 1-12. https://doi.org/10.1145/2010325.2010326

Fathy, Y., Barnaghi, P., & Tafazolli, R. (2018). An Online Adaptive Algorithm for Change Detection in Streaming Sensory Data. *IEEE Systems Journal, 13*(3), 2688-2699. https://doi.org/10.1109/JSYST.2018.2876461

Foote, J. (2000). Automatic audio segmentation using a measure of audio novelty. In *2000 ieee international conference on multimedia and expo. ICME2000. Proceedings.* (Vol. 1, pp. 452-455). IEEE. https://doi.org/10.1109/ICME.2000.869637

Foote, J. T., & Cooper, M. L. (2003). Media segmentation using self-similarity decomposition. In *Storage and Retrieval for Media Databases 2003* (Vol. 5021, pp. 167-175). International Society for Optics and Photonics. https://doi.org/10.1117/12.476302

Gharghabi, S., Yeh, C.C.M., Ding, Y., Ding, W., Hibbing, P., LaMunion, S., Kaplan, A., Crouter, S.E., & Keogh, E. (2019). Domain agnostic online semantic segmentation for multi-dimensional time series. *Data Mining and Knowledge Discovery, 33*(1), 96–130. https://doi.org/10.1007/s10618-018-0589-3

Gibb, B., Gibb, R., & Gibb, M. (1977). Stayin' alive. In *Saturday Night Fever, The Original Motion Picture Soundtrack*. Germany: RSO.

Gong, D., Medioni, G., & Zhao, X. (2014). Structured time series analysis for human action segmentation and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 36*(7), 1414–1427. https://doi.org/10.1109/TPAMI.2013.244

Kahol, K., Tripathi, P., & Panchanathan, S. (2004). Automated gesture segmentation from dance sequences. In *Sixth IEEE International Conference on Automatic Face and Gesture Recognition, 2004. Proceedings.* (pp. 883–888). IEEE. https://doi.org/10.1109/AFGR.2004.1301645

Krüger, B., Vögele, A., Willig, T., Yao, A., Klein, R., & Weber, A. (2016). Efficient unsupervised temporal segmentation of motion data. *IEEE Transactions on Multimedia, 19*(4), 797-812. https://doi.org/10.1109/TMM.2016.2635030

Krüger, V., Kragic, D., Ude, A., & Geib, C. (2007). The meaning of action: A review on action recognition and mapping. *Advanced robotics, 21*(13), 1473-1501. https://doi.org/10.1109/TMM.2016.2635030

Lin, J.F.S., Karg, M., & Kulić, D. (2016). Movement primitive segmentation for human motion modeling: A framework for analysis. I*EEE Transactions on Human-Machine Systems 46*(3), 325–339. https://doi.org/10.1109/THMS.2015.2493536

Liu, S., Yamada, M., Collier, N., & Sugiyama, M. (2013). Change-point detection in time-series data by relative density-ratio estimation. *Neural Networks, 43*, 72-83. https://doi.org/10.1016/j.neunet.2013.01.012

Markou, M., & Singh, S. (2003). Novelty detection: a review—part 1: statistical approaches. *Signal processing*, *83*(12), 2481-2497. https://doi.org/10.1016/j.sigpro.2003.07.018

Mendoza, J.I. (2014). Self-report measurement of segmentation, mimesis and perceived emotions in acousmatic electroacoustic music. Master's thesis. University of Jyväskylä. http://urn.fi/URN:NBN:fi:jyu-201406192112

Mendoza, J. I., & Thompson, M. (2017). Modelling Perceived Segmentation of Bodily Gestures Induced by Music. In *ESCOM 2017: Conference proceedings of the 25th Anniversary Edition of the European Society for the Cognitive Sciences of Music (ESCOM)*. Ghent University. http://urn.fi/URN:NBN:fi:jyu-201711024121

Otondo, F. (2008). Ciguri. In *Tutuguri*. Sargasso.

Patterson, T., Khan, N., McClean, S., Nugent, C., Zhang, S., Cleland, I., & Ni, Q. (2016). Sensor-based change detection for timely solicitation of user engagement. *IEEE Transactions on Mobile Computing*, *16*(10), 2889-2900. https://doi.org/10.1109/TMC.2016.2640959

Petzold, C. (ca. 1725). Minuet in G major. *The Anna Magdalena Bach Notebook*, Anh. 114.

Rodrigues, J., Probst, P., & Gamboa, H. (2021). TSSummarize: A Visual Strategy to Summarize Biosignals. In *2021 Seventh International conference on Bio Signals, Images, and Instrumentation (ICBSII)* (pp. 1-6). IEEE. https://doi.org/10.1109/ICBSII51839.2021.9445154

Schätti, G. (2007). Real-Time Audio Feature Analysis for Decklight3. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.85.7916&rep=rep1&type=pdf

Tardieu, D., Chessini, R., Dubois, J., Dupont, S., Hidot, S., Mazzarino, B., ... & Visentin, A. (2009). Video Navigation Tool: Application to browsing a database of dancers' performances. *on Multimodal Interfaces eNTERFACE'09*, 35. http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=0249E27EDBD8D12E8FF58DE4F9ABC18A?doi=10.1.1.159.3151&rep=rep1&type=pdf

Zacks, J. M., Kumar, S., Abrams, R. A., & Mehta, R. (2009). Using movement and intentions to understand human activity. *Cognition*, *112*(2), 201-216. https://doi.org/10.1016/j.cognition.2009.03.007

Zameni, M., Sadri, A., Ghafoori, Z., Moshtaghi, M., Salim, F. D., Leckie, C., & Ramamohanarao, K. (2020). Unsupervised online change point detection in high-dimensional time series. *Knowledge and Information Systems*, *62*(2), 719-750. https://doi.org/10.1007/s10115-019-01366-x

Zhou, F., De la Torre, F., & Hodgins, J. K. (2012). Hierarchical aligned cluster analysis for temporal clustering of human motion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(3), 582-596. https://doi.org/10.1109/TPAMI.2012.137

Authors' Note

This study was partially funded by the Finnish Foundation for Technology Promotion (Tekniikan edistämissäätiö). All correspondence should be addressed to Juan Ignacio Mendoza, at the University of Jyväskylä, email: juigmend@student.jyu.fi

Human Technology
ISSN 1795-6889
https://ht.csr-pub.eu