Dipartimento di Informatica, Bioingegneria,
Robotica ed Ingegneria dei Sistemi

CEST: a Cognitive Event based Semi-automatic Technique for behavior segmentation

by

Eleonora Ceccaldi

Theses Series DIBRIS-TH-2021-

DIBRIS, Università di Genova

Via Opera Pia, 13 16145 Genova, Italy http://www.dibris.unige.it/

Università degli Studi di Genova

Dipartimento di Informatica, Bioingegneria,

Robotica ed Ingegneria dei Sistemi

Ph.D. Thesis in Computer Science and Systems Engineering
Computer Science Curriculum

CEST: a Cognitive Event based Semi-automatic Technique for behavior segmentation

by

Eleonora Ceccaldi

May, 2021

Dottorato di Ricerca in Informatica ed Ingegneria dei Sistemi
Indirizzo Informatica
Dipartimento di Informatica, Bioingegneria, Robotica ed Ingegneria dei Sistemi
Università degli Studi di Genova

DIBRIS, Univ. di Genova
Via Opera Pia, 13, I-16145 Genova, Italy
`http://www.dibris.unige.it/`

Ph.D. Thesis in Computer Science and Systems Engineering
Computer Science Curriculum
(S.S.D. INF/01)

Submitted by Eleonora Ceccaldi
DIBRIS, Univ. di Genova
`ceccaldi.infomus@gmail.com`

Date of submission: March 2021

Title: CEST: a Cognitive Event based Semi-automatic Technique for behavior segmentation

Advisor: Gualtiero Volpe
Dipartimento di Informatica, Bioingegneria, Robotica ed Ingegneria dei Sistemi
Università di Genova
`gualtiero.volpe@unige.it`

Ext. Reviewers:
Benoît Bardy
EuroMov
Università di Montpellier
`benoit.bardy@univ-montp1.fr.`

Rossana Damiano
Dipartimento di Informatica
Università degli studi di Torino
`rossana@di.unito.it`

Alessandro D'Ausilio
Dipartimento di Scienze Biomediche e Chirurgico Specialistiche
Università degli studi di Ferrara
`dsllsn@unife.it`

Abstract

This work introduces CEST, a Cognitive Event based Semiautomatic Technique for behavior segmentation. The technique was inspired by an everyday cognitive process. Humans make sense of what happens to them by breaking the continuous stream of activity into smaller units, through a process known as segmentation. A cognitive theory, the Event Segmentation Theory, provides a computational and neurophysiological account of this process, describing how the detection of changes in the current situation drive boundary perception. CEST was designed to provide affective researchers with a tool to semi-automatically segment behavior. Researchers investigating behavior, as a matter of fact, often need to parse their research data into simpler units, either manually or automatically. To perform segmentation, the technique combines manual annotations and the output of change-point detection algorithms, techniques from time-series research that afford the detection of abrupt changes in time-series. CEST is inherently multidisciplinary: it is, to the best of our knowledge, the first attempt to adopt a cognitive science perspective on the issue of (semi) automatic behavior segmentation. CEST is a general-purpose technique, as it aims at providing a tool for segmenting behavior across research areas. In this manuscript, we detail the theories behind the design of CEST and the results of two experimental studies aimed at assessing the feasibility of the approach on both single and group scenarios. Most importantly, we present the results of the evaluation of CEST on a data-set of dance performances. We explore seven different techniques for change-point detection that could be leveraged to achieve semi-automatic segmentation through CEST and illustrate how two different bayesian algorithms led to the highest scores. Upon selecting the best algorithms, we measured the effect of the temporal grain of the analysis on the performance. Overall, our results support the idea of a semiautomatic segmentation technique for behavior segmentation. The output of the analysis mirrors cognitive science research on segmentation and event structure perception. The work also tackles new challenges that may arise from our approach.

To all my safe places.

Winds in the east, mist coming in. Like somethin' is brewin' and about to begin. Can't put my finger on what lies in store, but I fear what's to happen all happened before." (Mary Poppins)

Acknowledgements

I wish to thank all the people that made this thesis possible. I am humbled by the amazing and inspiring reviews I received, that helped me submit the best possible version of this work. I feel very lucky for all the kindness and encouragement I received from Antonio and all the staff of CasaPaganini, current and former. The biggest ”thank you” should go to Simone and Roberto, for answering all my questions and requests, even the most annoying ones. Last but not least, I am grateful for Rossana, that never stopped mentoring me, and for Gualtiero, by far the most humble, patient and encouraging person I’ve ever met.

Table of Contents

I Introduction

1 General outline of the work

This thesis describes the outcome of my PhD course in Computer Science and Systems Engineering at University of Genoa. The work, carried out with Professor Gualtiero Volpe, is aimed at proposing a novel semi-automatic technique for behavior segmentation. The technique has two core characteristics: it was designed with a highly-interdisciplinary approach and it is a general purpose technique, that researchers can adapt to their specific research questions. To reach such ambitious goals, we tried our best to combine a cognitive science approach to segmentation and most up-to-date techniques for automatic change detection from computer science. This manuscript illustrates all the steps of the process that led us to our final semi-automated technique. This section details the motivations of the work and gives a glossary that might help the reader avoid confusion. Section II illustrates the topic of segmentation, from how it was theoretically described in cognitive science to how the problem is dealt with in affective computing and computer science. Section III illustrates the feasibility studies we ran when initially designing our technique. The study are described more thoroughly in [19] and [20]. Sections IV and V are the core of the work. The former describes the general functioning of the technique, while the latter reports the results of a set of experiments that where conducted for the design of the technique, in terms of parameters, modules, etc. and for testing its performance against ground truth data. Section VI ends the manuscript.

2 Motivation and goals

This work is motivated by the idea that interdisciplinary research can be the key to address complex research issues. A very common issue in affective computing research, the partitioning of data into smaller units has been theoretically addressed by cognitive science as well, as breaking the reality into smaller units is also something human beings do to give meaning to what they perceive [105] and it is something they do from a very young age [8]. As a consequence, this thesis aims at bridging this gap, giving a cognitive science solution to a computer science problem.

The goal of the work is therefore the creation of a semi-automatic technique for movement segmentation. The technique has its theoretical foundation in how the process of perceiving the structure of the ongoing situation happens in the human mind, as described by the Event Segmentation Theory [108]. The technique we propose is general-purpose, meaning it can be adapted to different research requirements. We think adopting a multi-disciplinary approach may have helped us reach this goal. However, we live up to the reader to decide whether we succeeded or not.

3 What's in a name?

Segments, clips, units, time-windows, fragments: many terms can be found referring to the results of breaking a long sequence into smaller, simpler units. Section II will attempt at clarifying all the different views, approaches and terminologies to the topic. Here, we report the terms that will be used throughout this manuscript, with the goal of helping the reader go through the sections more smoothly. Table 1 also includes definitions and references from relevant research.

Word	Meaning	References
segmentation	the process of breaking the material into smaller portions	[104]
unitizing	the segmentation of data prior annotation or labeling	[19]
cognitive segmentation	the human perception of beginnings and endings of events	[108]
event	a portion of time perceived by an observer as having a beginning and an end; psychological events have durations on a human scale, spanning from a few seconds to tens of minutes	[108], [104]
change categories	the different kinds of changes that correlate with the probability of perceiving a boundary in a situation	[108]
boundaries	the beginnings and ends of segments	[108]
change-point detection CPD	the automatic recognition of abrupt changes in a time-series	[6]

Table 1: Terms and definitions used in this manuscript

II State of the art: from event segmentation to unitizing

This section illustrates the issue of parsing a continuous stream of activity from different points of view. Each paragraph deals with the topic from a different theoretical perspective and shows the theories and methodologies that have been proposed. The discussion begins with cognitive segmentation, detailing how cognitive science has described this process as it unfolds in the human mind. It continues by describing how this issue is dealt with in affective computing, when the segmentation is a tool, when not a necessary step, to observe human behavior and interaction. Then, it moves on to change point detection, sketching the most common approaches to the automatic detection of segment boundaries in time-series. Finally, it provides a framework for evaluating segmentation algorithms from movement segmentation research, and concludes by illustrating the main areas of novelty of the technique we propose.

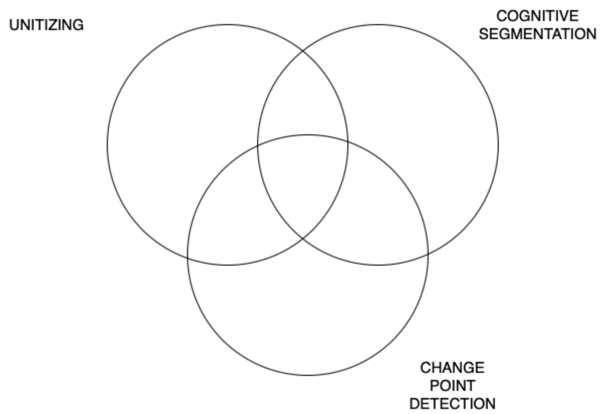


Figure 1:
Three different perspectives on segmentation.

Figure 1 depicts the connections among the three different perspectives. Cognitive segmentation refers to the cognitive process of perceiving the continuous flux of experience as composed of discrete events. As for the unitizing that is performed in affective computing research, the target of cognitive segmentation is usually human behavior, whether in the form of mere motion or social interaction. Similarly to change-point detection, segmentation is performed by leveraging changes in the situation as cues for detecting boundaries in the stream of activity. Whereas affective computing unitizing and change-point detection come from research purposes, cognitive segmentation is a general everyday process, that process through which humans make sense of what happens to them [105].

1 Segmentation in cognitive science: the Event Segmentation Theory

Figure 2:
Two groups of black spots also known as dalmatian dogs

Figure 3 shows two groups of black spots. However, the authors are confident in assuming that the reader immediately saw two dogs in it. This may have happened because the reader has a dog at home or knows such a dog, or because, as a kid, he or she watched 101 Dalmatians many times. In other words, the reader holds, in her mind, the concept of dog. In fact, object perception works through concepts: concepts from past experience or knowledge, stored in long-term memory, drive the understanding of the world. Humans, in the words of [94], possess *the seemingly effortless ability to perceive meaningful object*. As a consequence, the world is not perceived as made up of a continuous stream of colors, but rather as made up of discrete entities. The same holds in time: the world is not experienced as an endless stream of activity but it is perceived as a series of discrete events. Similarly to concepts, or objects, cognitive representation of events guide our perception of time. Such representations are called *event models*. For instance, putting hot water into a cup, placing a tea-bag in it, and waiting for a few minutes are grouped into a "making tea" event model. Event models are not fixed: they can be discarded or updated when no longer suitable to describe the current state of affairs. The updating of event models depends on changes in the situation. In the "tea" example, one might predict sugar being added to the tea. This "fortune telling" function of event models lasts until it is no longer possible to predict the future unfolding of events by relying on the same representation. A jumping-jack performed after the tea has been drunk might be difficult to integrate into the "making tea" event. This might indicate the need of reshaping the model, or else it might imply that the "making tea" event has ended and a new "doing gymnastics" event has begun.
Although intuitively the world presents itself to human beings as a seamless stream of ongoing perception, the phenomenological experience is that of discrete events. Research on the temporality of conscious experience has extensively dealt with this topic, debating over the continuous

or discrete nature of consciousness [33] [96]. Humans are directly aware of the succession of brief temporal intervals, i.e. events over time[29]. The Ecological Theory of Event Perception [89] has posited that events are in fact the sources from which perceptual knowledge arises. Building from this, the Theory of Event coding [43] has claimed that events, defined as cognitive representations of *bundles of features* underpin both perception and action planning. *Event structure perception* [105] is an automatic component of human perceptual processing. This seems to be a very early-emerging skill: research showed that even 10 months old infants can perceive action boundaries. When presented with video-taped actions cut before the ending of the sequence, children paid more attention, as if expecting to the ending of the sequence they had predicted [42]. Studies, e.g. [76], have demonstrated that the perceptual organization of ongoing behavior can be measured: by asking participants to watch a video-clip depicting daily life activities and to press a key whenever they feel *"a meaningful portion of action ends and another one begins"*. When naive participants are asked to place boundaries in a scene, such boundaries are reliable across viewers and over time.
The Event Segmentation Theory (EST) [108] has described the innate ability of human beings to parse an ongoing interaction into meaningful units and provides a computational and neurophysiological account of event structure perception. Zacks [105] defines an event as *"a segment of time at a given location, that is conceived by an observer to have a beginning and an end"*. The core of the theory lies in three main ideas:

- event segmentation is an automatic, ongoing component of human perception
- segmentation that occurs during perception scaffolds later memory recall and learning
- event boundaries are localized by identifying meaningful changes in physical and social features.

Event segmentation relies on event models observers form of the ongoing situation. Event models are based on perception and previous experience. Such models frame new incoming information and guide prediction of future developments. In a sense, event models are the result of statistical structure learning [9]. What is more, the perception of event boundaries is closely tied to prediction: a boundary is perceived whenever unpredictable changes in salient features occur, putting the currently active event model at stake.

Boundary perception is also highly correlated with movement perception: brain regions that have been found more active during boundary perception are those that are known to be involved in movement perception [92]. Movement perception is surely a part of the story of boundary perception, but the process does not end there. If movement perception is tightly linked with fine-grained segmentation, when segmentation gets more coarse-grained actors' goals and cause-effect relations play a very important role as well [108]. Event segmentation is indeed achieved

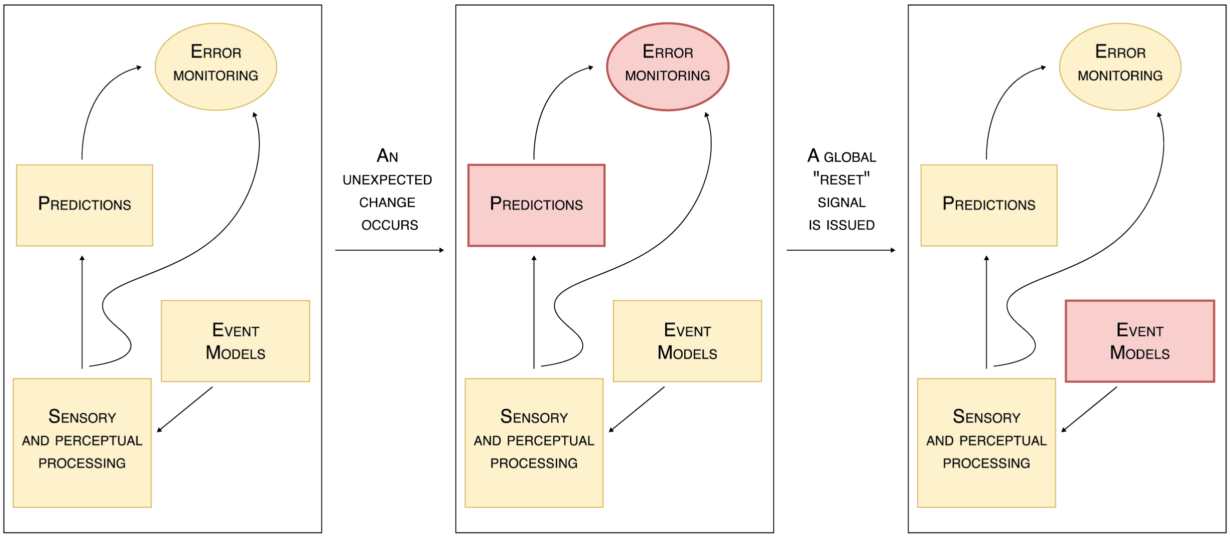


Figure 3:
A sketch of the EST, adapted from [59]. (I) Sensory and perceptual processes leverage event models to form predictions. When predictions fail and an unexpected change occurs, the error monitoring process detects an error. This detection triggers an ”error signal” that can lead the current model to be readjusted or replaced. In the latter case, a boundary is perceived between the current event and the following, now correctly described by the new model.

through a combination of bottom-up and top-down processes. The perception of bottom-up features, such as kinematics, is integrated with and interpreted through top-down information from event models.
Boundary perception leverage information from several different situational dimensions. A set of studies in [104] shed light on how situational cues affect event segmentation. Observers seemed to analyze the situation by attending to several different dimensions at once, and by tracking changes in such dimensions. In these studies, behavioral data demonstrated the association between changes and boundary perception. Boundaries were identified as a consequence of changes in 7 dimensions: (1) *time*, (2) *space*, (3) *objects*, (4) *characters*, (5) *character interaction*, (6) *causes*, (7) *goals*. Moreover, effect is incremental: the more situational features change, the larger the probability that a viewer would identify an event boundary.
Recently, some revisions to the EST have been proposed. One major novelty regards boundaries: according to [58], boundaries are not detected, they are predicted. Another addition to the theory regards the importance of conceptual features, such as cause-effects relationships and goals, in placing boundaries: segmentation is, as mentioned before, achieved by leveraging knowledge on the statistical structure of events [57]. Mostly, current research says, statistical regularities are learned, as the mind develops, in the shape of goal structures [62], [63].
To sum up, a continuous stream of activity is perceived as discrete events, through the perception of boundaries between events; such anticipated perception is driven by changes in the situa-

tion, that serve as indicators of a mismatch between the statistical cognitive representation of the event that is available from previous learning and the current situation at hand. This everyday and spontaneous perceptual process is also a research step in many research areas. Instances of such applications will be presented in the following section.

2 The issue of unitizing: state of the art

This section section will focus on the segmentation of behavior streams in the specific area of application of affective research, as it exemplifies the complexity of applying this spontaneous and effortless cognitive process to practical research problems. The book by Picard [81] forged the field of *affective computing* (here, AC), giving rise to a broad, interdisciplinary area of research on emotion, interaction and, more generally, human behavior. It could be argued that this new approach to the observation, recording and analysis of behavior came with the need for methods to segment such behavior. When conducting behavior research, the decision on the segmentation approach might be needed from the very beginning, when a data-set is created and it needs to be organized and stored, as in [109]. Moreover, the creation of data-sets and corpora is strictly linked with the manual or automatic annotation of the behaviors or the affective states the corpora focus on. In this phase, behaviors can be annotated continuously or they can be segmented prior annotation. What is more, the aforementioned annotations are often performed by providing the annotators with coding schemas, that guide the observation and coding of behavior. When adopting a specific coding scheme, again a decision needs to be made on the temporal granularity of the ratings. Regarding this, Meinecke and colleagues [74] suggest answering the following question: are behavioral codes assigned to a behavioral event or are codes assigned to a specific time interval? To sum up, segmentation might be needed after data-collection, before annotation or during the selection of a coding scheme.
Despite being a common ”sword of Damocles” for researchers, no work, to the best of our knowledge, has attempted at unifying the different approaches to the issue. To shed more clarity and gather knowledge on the variety of approaches to segmentation that can be found in AC, we went through all the papers ever published in the highest-ranking journal in the field, according to scimagojr.com, IEEE Transaction on Affective Computing [1] and by reading related articles. From the aforementioned sources, we selected and analyzed papers meeting the following criteria:

- the study regards the analysis of behavior recorded through video camera or motion-capture technologies;
- the study or data-collection described in the paper contains annotation or labeling;

[1]https://ieeexplore.ieee.org

- the rationale behind segmentation, or the decision to opt for continuous annotation, is explicitly reported in the paper.

Papers where the unitizing, and hence the analysis, was only performed on audio features were excluded from our studies, as speech-based segmentation goes beyond the scope of our work. Similarly, we weren't interested in analyzing language-based segmentation, or segmentation of narratives. The reader might explore these topics in [90], [78] or [30]. It is worth mentioning, however, how interesting parallelisms can be found between event-based unitizing and floor control shift detection. In this area of research, interactions are temporally analyzed against two different levels: at a finer level, considering sentence units, and at a coarser level, focusing on floor control change (i.e. change of the speaker). In both cases, the duration of the units is variable, depending on the interaction itself [26].
For, us the quest for clarity on the issue began by searching how different research name the issue of parsing their material. In fact, the process of breaking the material into smaller units has not received a common definition yet, often resulting in theoretical ambiguity. The meaning of the term "segmentation" varies across different fields. In marketing, it refers to *the process of dividing a heterogeneous market into relatively more homogeneous segments*[2]. In computer vision, it indicates *the process of partitioning a digital image into multiple segments (sets of pixels, also known as image objects)*[3]. Out of the total number of papers that were analyzed for this survey, most referred to the process of making the material shorter by using the term *segmentation*, whereas a smaller portion referred to unitizing as *slicing*, or *clipping*. When the unitizing is performed as a prior step towards labeling, often the term *time-window* is used to define the output of such process. As mentioned in section I, we use *segmentation* as a general term to refer to the process of breaking the reality into smaller units and *unitizing* to indicate segmentation for research purposes.
Committing to a specific unitizing or segmentation approach can be necessary in different research steps. It can be needed upon the organization of a data-set, for example, to parse long recordings into meaningful events, usually by manually segmenting specific behaviors of interest, or it can be necessary to make video shorter, for instance by selecting a fixed time-window and by automatically cutting the videos accordingly. Segmentation can be needed after the creation of a data-set, for instance, to utilize the recordings for a specific research purpose, such as the analysis of specific turn-taking configurations, a pattern of dance moves, and so on. Moreover, when a coding scheme is leveraged for the annotation or labeling step, a decision on the temporal grain has to be made, to choose between *interval coding* [97], standing for identifying a fixed-length interval of time for raters to note the occurrence of any of target behavior, *continuous coding* meaning the non-stop annotation of the behavior or interaction or *event-based coding*, when specific portions of the scenario of interest are isolated, based on the specific research question at hand. Interestingly, in some cases annotation can coincide with segmentation. In [87], the authors asked their participants to annotate disrespectful behavior in videos, by highlighting

[2]https://en.wikipedia.org/wiki/Marketsegmentation
[3]https://en.wikipedia.org/wiki/Imagesegmentation

the portions of time where the behavior could be observed. In other words, the participants annotated the interactions by segmenting them. However, usually researchers commit to one of the possible approaches, each coming with its own shortcomings and virtues.
When a fixed-length time-window is selected for unitizing or coding, researchers may opt for a time span suggested by psychology research on the time spans of behavioral observation. Ambady illustrated how thin-slices of behavior can be enough to gather meaningful observations [5], [3]. Humans are able to make accurate enough inferences on others' personality traits [11], sexual orientation [4], racial bias [84] intelligence [11] and academic achievement [2] after observing them for less than 1 minute. These perceptual studies had a great impact on affective computing, paving the way towards the annotation, labeling or automatic detection of target behaviors from small portions of time. Gatica Perez and colleagues used this approach to detect high interest level in meetings [38]. Hung and colleagues for the first time adopted thin slices for automated estimation of cohesion [44].
The selection of the window-size usually depends on the specific research question at hand. In their work on emotion recognition, [82] trim their data into less than 500ms and 500ms to 4s clips, as, in their view, the two different durations are needed in order to distinguish between macro and micro emotion clips. To investigate the possibility of annotating through an iterative approach, [35] adopted 20s long segments.
Whitehill and colleagues adopt 10s long segments for their study on engagement detection in students [102]. From our research on papers published in affective computing journals and conferences, we noticed that when a fixed-length window is selected, the rationale behind the selection is seldom reported in the paper. In [102], however, the authors motivate their choice by illustrating how, in the context of engagement observation, shorter clips do not provide enough context and inter-rater reliability is affected, whereas with longer clips tend to be harder to evaluate because they may mix different levels of engagement. In this, we believe, lies one of the main shortcomings of fixed-window segmentation: selecting the most appropriate duration is not trivial as it can easily affect the results, as we demonstrated in [19] and [20]. It must be noted, however, how fixed-window unitizing comes with many advantages: it can be easily automatized, it is less prone to researchers' subjectivity and it requires no training.
The work of [65] illustrated the shortcomings of fixed-length windows, in their case 1 second long, for fusing different modalities for event-based affect recognition. The authors demonstrate how a fixed-window approach can be challenging when multiple modalities need to be taken into account, as the features that are needed for affect recognition can have different time-spans. As a consequence, they advocate for an event-based approach. Similarly, AC research has shown how affective phenomena, either natural [98] or artificial [77], have an inherent temporal sequencing that should not be neglected, as it may be the case with a fixed-window segmentation. Such an approach, in fact, might lead to boundaries of the units being placed within actions, thus resulting in losing important information. For example, it might split an interaction before its ending (e.g., think of 2 people interrupted as they speak or of a smile following a seemingly harsh statement: it might overturn the meaning of it, but would the statement and the smile end up into different segments, such overturning will be lost).

In event-based or variable-windows approaches, the window changes according to the specific needs of the study. Often, this kind of segmentation coincides with manual segmentation, for instance when it is aimed at isolating specific behaviors of interest. [28], in their study on fake smiles, performed a temporal segmentation on their data, based on the identification of peak-frames, i.e. frames portraying most strongly expressed smiles, and the subsequent creation of fixed-length segments containing such frames. [22] manually parsed their human-robot interactions into question-answer segments, to isolate the phenomenon of interest in their study. It must be noted, however, that variable-length, data-driven segmentation does not necessarily coincide with manual segmentation. Hemamou and colleagues [41] proposed an automatic methodology for identifying slices of attention in job interviews. Their approach is automatic, yet it adopts variable-length time-windows.

Our technique hopes to provide a solution that is both easier and less time-consuming than fully-manual annotation, while also avoiding abrupt cuts that can put the integrity of annotations or labeling at risk.

Outside the field of affective computing, time-series research has also provided many algorithms that segment by leveraging change points in a data-stream, in a way that gives changes the same importance they hold in cognitive segmentation. This field is known with the name of *change point detection*.

3 From unitizing to change point detection

The theory that describes cognitive segmentation, the Event Segmentation Theory [107], highlights the role of changes in driving the perception of boundaries. As the EST shows, boundaries are perceived by human observers when changes occur. As a consequence, it could be argued that segmentation is, at its core, boundary detection, or at list strongly connected to it. Change-point detection (here, CPD) or analysis consists of identifying the inference of a change in distribution for a set of time-ordered observations [72]. CPD means, according to [14], *partitioning an input time series into a number of segments*. In general, the goal of change-point detection is to detect whether and when abrupt distributional changes take place in a time series [25]. It is a multidisciplinary area of research and it is crucial across many fields such as climate science, finance, signal analysis, and so on. Many works have applied CPD to behavior detection or monitoring [23]. As pointed out in [6], comparing the performance of different CPD algorithms is difficult, as the algorithms are usually tested against data-sets that are very different. Also, CPD algorithms are seldom evaluated on real-world data [14]. Although the field started by dealing with univariate time-series, with techniques focusing on mean, variance and autocorrelation [15], current CPD research concentrates also on multivariate data. When dealing with multivariate data, the issue of identifying change points requires taking into account the changes in the different variables as well as the changes in the relationships among them. For instance, in emotion research bodily, cognitive and environmental changes are observed, as well as changes in the correlation among such factors [15]. Change-point detection is closely linked to change-point estimation. Differently from CPD, however, change point estimation aims to model and interpret known changes in time series rather than detecting a change that has occurred. Non-parametric multivariate CPD algorithms have been proposed, detecting changes in correlations and in means. However, comparing such algorithms is often hard, as they are based on different statistical approaches. Cabrieto and colleagues, therefore conclude how the decision on the algorithm heavily depends on the data setting at hand, and how adopting different methods can help address all the different changes involved.

Change-point detection algorithms can be divided into "online" and "offline" or "retrospective". Offline algorithms consider the entire data set at once, and then detect changes *a-posteriori*. Generally all change points from a given time sequence are identified in batch mode. On the other hand, online, or real-time, algorithms run concurrently with the process they are monitoring, processing each data point as it becomes available, aiming to detect a change-point as soon as possible, ideally before the next change occurs [6]. Some argue that the results of offline CPD algorithms are more robust [66]. A thorough survey of available CPD algorithms goes beyond the scope of the current paper. The reader could understand more about the topic by reading the survey by [51] and [6]. This paper will focus on a set of offline algorithms, that are, in our opinion, viable options to achieve the automatic cognitive-inspired segmentation this work aims at. In section V we present different CPD algorithms that can be exploited for behavior segmentation.

It is worth mentioning that other possible automatic segmentation approaches are available from research, for instance from computer vision literature. Such approaches, however, differ from our perspectives and objectives and, therefore, go beyond the scope of our work. To have a more thorough understanding of the topic, the reader might search into the computer vision literature. More specifically, [100] offer a thorough survey on action recognition and segmentation methodologies, that can help understanding how computer vision has approached the issue of boundary identification.

4 Movement primitives segmentation

The issue of identifying starting and ending points of segments is not exclusively related to affective computing and wasn't only tackled through change-point detection. In movement analysis, it often coincides with the need to parse movements into simpler action sequences, for instance by relying on movement libraries [73], [10]. The term *movement primitives segmentation* indicates the segmentation of movement data into smaller components and the identification of segment points, i.e. starting and ending of each segment [64]. Generally, this kind of segmentation is aimed at facilitating movement recognition, modeling and learning. In [64], Lin and colleagues provide a framework for comparing different segmentation algorithms. Although their segmentation approach was proposed in the field of movement segmentation, we believe their framework to be useful for segmentation algorithms in general, as it highlights several crucial aspects of segmentation algorithms. Five different components can be leveraged to characterize any segmentation algorithm:

- **segment definition:** this aspect indicates how segment boundaries are characterized in the algorithm. Boundaries can depend on inner movement characteristics, usually domain-specific (e.g. a punch), can be defined through changes in metrics from the data-stream (e.g. changes in variance) or can be determined *a-priori*, for instance through templates.

- **data collection:** algorithms can work on data from various sources. For movement analysis, motion capture technologies and cameras are the most common and fruitful approaches, but other sensors and data sources can be applied, such as Inertial Measurment Units (IMUs). This algorithm characteristic also regards the possibilities for collecting the ground-truth, that can be obtained by having human observers label the data or by having them directly segment the data stream to identify starting and ending points of segments.

- **application-specific requirements:** segmentation algorithms may be distinguished for their ability to perform on data different from the training set. Often, however, they can be very domain-specific.

- **design:** design criteria shape the algorithm. After the decision on the previously mentioned aspects has been made, the design process continues with the preprocessing that is performed on the data, such as the application of low-pass filters to remove high-frequency noise or the transformation of the feature space, with or without dimensionality reduction. Also, a decision has to be made on the windowing (i.e., fixed vs. variable-length windows).
  Algorithms can be clustered into a) *online segmentation approaches*, b) *semionline segmentation approaches*, c) *online supervised segmentation approaches*, d) *online unsupervised segmentation approaches*.

- **verification:** a verification technique is selected to evaluate the performance of the algorithm against a specific ground-truth.

Although, as mentioned before, this classification was proposed specifically in the field of movement segmentation, it highlights common approaches and challenges that can help understand the problem of segmenting through algorithms in general. In fact, a common issue seems to be the possibility to achieve effective segmentation despite the specific research domain. We hope our contribution can pave the way towards general-purpose segmentation algorithms.

5 Main contributions to the field

Our work was heavily inspired by the Event Segmentation Theory [108] and by its most recent revisions [63]. Nonetheless, our contribution aims at moving from the theory, leveraging it to create a tangible tool for researchers. Although we aim at proposing a general-purpose technique, we envision its major area of application in the field of affective computing. In affective computing, in fact, often a segmentation step is needed prior to annotation or labeling. Moreover, research in affective computing tackles complex phenomena, such as emotional behavior and social interaction, whose investigation requires assessing both low-level features, like kinematics and high-level features such as cause-effects relations and goals. In this, our work adopts a novel approach, as segmentation is often performed manually or automatically by adopting a fixed window. Our technique, instead, performs a semi-automatic segmentation that follows how cognitive segmentation is achieved. Our proposal shares with change-point detection from time-series analysis the centrality of changes, although it stands out as segmentation is achieved through the theory-compliant combination of the outputs of parallel change point detection from different time-series. Our technique matches the importance of low-level kinematics with movement primitives segmentation, although, in our case, low-level, mid-level and high-level features are combined.

III Preliminary studies

Before testing our automatic technique, we explored the feasibility of a EST segmentation approach through a set of experimental studies. The approach was tested against single user scenarios in [20] and social interactions in [19]. In both studies, unitizing was manually performed according to the following steps:

- *Step 1 – Annotating changes*: annotating each change in the scene that can be related to one of the 7 categories. The first step would be parsing the material and keeping track of each change category that can be distinguished.
- *Step 2 – Placing boundaries*: detecting boundaries based on the changes in the scene. According to [104], changes in different categories correlate with boundary perception, with a spike in the probability of a boundary being detected after three changes. For this reason, we cut the scene after changes in three different categories are found. What is more, Levine and colleagues [61] illustrate how goals elicit boundary perception more than other lower-level changes. As a consequence, in our technique, changes belonging to the "goal" category value twice.

The next paragraphs describe such studies more thoroughly.

1 Testing the approach in a single user context

In [20], we assessed our technique on videos depicting single users. To this aim, we compared our approach with one of the most commonly used approaches in automatic video analysis, i.e. segmenting the video in fixed-length windows. Comparison was performed with respect to the effect of the unitizing approach on annotation of video material. The study demonstrated that cognitive-based segmentation better reflects the variance of the raters' scores and it represents, in the single-user context as well as in social interaction, a viable option for unitizing for behavior analysis. In the study, we opted for a domain, i.e. public speaking performances, that could afford comparison with the results obtained in [19] and described in subsection 2. Hence, we decided to examine communication in terms of the *language functions* conveyed during the performances in order to have the material annotated in terms of social and non-social aspects, as in our cohesion studies described further.
According to [45], each language unit (e.g. a sentence or a word) can be classified according to its contribution to the communication between the sender and the receiver of the conveyed message. More precisely, six different functions of language can be distinguished:

- referential: describing a situation, object or mental state

- poetic: focusing on the message *per se*, like in poetry
- emotive: adding information on the addresser's mental state
- conative: directly engaging the addressee, as in imperatives
- phatic: affording, initiating and maintaining interaction, for instance in greetings
- metalingual: reflecting and communicating about the language itself.

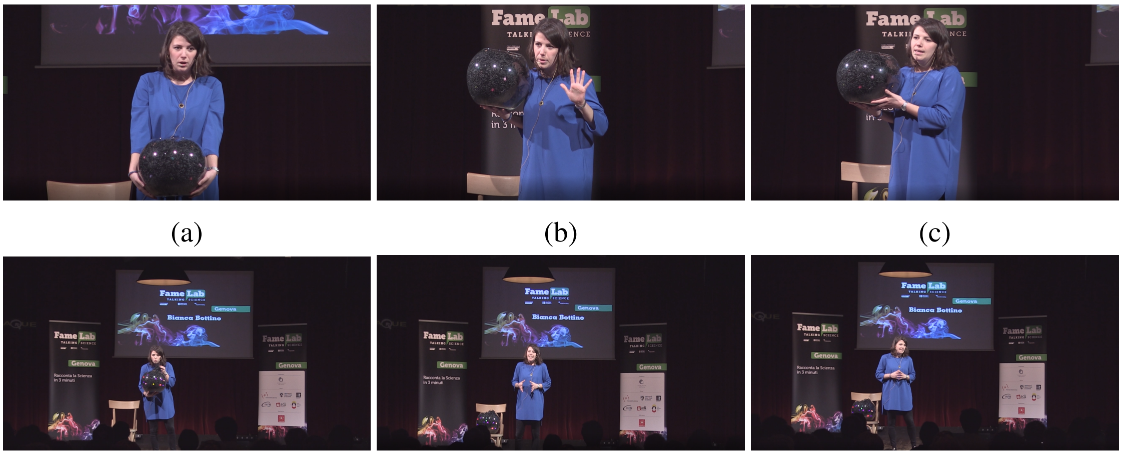
Thoroughly describing such classification of the functions of language goes beyond the scope of this manuscript. However, it is worth noting how some of these functions can be clustered into a *task* and *social* oriented dimension. In fact, the referential and poetic functions serve the purpose of properly conveying the intended message, whereas the phatic and conative functions address the interactive and social aspects of communication.
After identifying our testing field, we defined how the EST-based approach could be operationalized to unitize our material. Table 8 illustrates how we operationalized each change category. Figure 4 further depicts our approach.

Changes in EST	**Changes in our EST-inspired technique**	**Example in public speaking**
C1.Time	Timing and rhythm of the interaction	The speaker starts gesticulating fast
C2.Space	Motion direction	The speaker starts pointing at something
C3.Objects	Interaction with objects	The speaker dismisses a tool she was using
C4.Characters	Character location	The speaker leaves the stage
C5.Character interaction	Interaction patterns	The speaker directly addresses the audience
C6.Causes	Causes and appraisal	Something happens as a consequence of a new state of affairs
C7.Goals	Goals fulfilled, dismissed, or replaced	The speaker terminates to illustrate a concept and moves to another

Table 2: The left column reports the changes categories as described in [104]. The column in the middle shows how these changes were operationalized in our EST-based technique. The right column provides an example of occurrence of each change.

In this experiment, we compared our approach with the common-practice thin-slices approach with respect to: a) the average agreement on the raters' scores; b) the extent to which these scores reflect the intrinsic variability of the data-set. To this aim, a pool of 35 external observers[4] rated coding units obtained by applying the two different unitizing techniques to a set of 10 videos of

[4]our participants were all volunteers, recruited online



(a) (b) (c)

(d) (e) (f)

Figure 4: Example of EST unitizing. The only changes from frame (a) to (b) and from frame (b) to (c) are in the velocity of the speaker's (hand) movements. The changes all fall into the same category, i.e., timing (C1), therefore the frames all belong to the same event. Instead, from frame (c) to frame (d) three different categories of changes are observed: again, in the timing of movements, but also in the interaction with the object (C3), as the speaker is now incorporating it in her talk, and in the goal of the speaker (C7), as she moves from introducing her talk to the core of her theme. According to our technique, frame (c) and (d) thus belong to different events. Differently, only one type of change (movement) is observed between frame (d) and (e) and between (e) and (f), all belonging to the same event.

public speaking performances. The stimuli used in this study are audio-video recordings from the FameLab public speaking competition held in Genoa, Italy. This data-set was selected as it contains diverse performances and is rich in subtle non-verbal cues. The videos portray FameLab finalists attempting to illustrate a scientific concept of their choice as clearly as possible in three minutes, while also engaging the audience. While doing this, they could use objects (e.g. a puppet) to enrich their presentation. Videos were 3 minutes long.
A trained psychologist watched each performance and unitized it according to the principles of the Event Segmentation Theory. Fig. 5 displays how changes were used to perform unitizing. Following the theory, a new coding unit is created whenever 3 changes are detected. Goal changes were scored double, due to the importance of goals in boundary perception posited by [62]. To compare the EST unitizing with a fixed-window automatic approach, we decided to automatically parse the videos into 20 seconds long segments (here, "AUT" segments), following [44].
To compare the effects of different unitizing techniques on the annotations, we devised a questionnaire assessing the language functions fulfilled by the speakers' performances, in terms of clarity and straightforwardness of the explanations and in terms of their ability to entertain and engage the audience. The former were assessed through *task* items, the latter through *social*

items (see table 3). The reason behind this is to map our results with our work on cohesion, which includes a task and a social component [18]. In this study, we adopted 5-points Likert items, from 1 (*Not at all*) to 5 (*Yes, definitely*) [31]. The questionnaire was in Italian. Ratings were collected via a web application in an anonymous form. First, raters were welcome, and given the instructions for their task. Then, they were asked to complete information about age and gender and informed on data protection policies. Finally, they started to watch the coding units, randomly administered, to be rated according to the items in Table 3. Each participant was presented with 20 units, with the whole test lasting approximately 15 minutes. However, raters could leave the experiment when they wished. The total number of gathered ratings was 495.

Item-task	**Item-social**
Illustrates the topic clearly	Makes jokes
Illustrates the topic thoroughly	Engages the audience
Helps the audience understand the topic	Amuses the audience
Describes the topic extensively	Interacts with the audience
Has the audience focus on the topic	Fosters audience participation

Table 3: Items in the questionnaire for the task and social oriented aspects of communication (English translation from Italian).

Unitizing	**ICC and 95%-CI**	
	Task	**Social**
AUT	0.98 (95% CI [0.97, 0.99])	0.96 (95% CI [0.94, 0.98])
EST	0.96 (95% CI [0.94, 0.98])	0.97 (95% CI [0.95, 0.98])

Table 4: ICC values and Confidence Intervals

To test the feasibility of our approach, we investigated the effect of the adopted technique in terms of agreement of raters on the scores given to the questionnaire and in terms of variance of such scores. Our analysis follows what illustrated in [19], where a similar technique was tested against social cohesion.
Inter-raters reliability was assessed[5] by means of a one-way average-measures (ICC) to evaluate whether raters agreed in their ratings across unitizing techniques. Table 7 shows the obtained values of ICC and the respective 95%-confidence intervals. All the resulting ICCs were in the *excellent* range [56], indicating that raters had a high degree of agreement and that both the task and the social dimensions were similarly rated across raters independently of the unitizing technique. Both unitizing techniques can be therefore considered viable to achieve suitable ratings. For further analysis, the average of the raters' scores is assigned to each coding unit.

[5]The significance threshold for all the tests in this study was set at .05

To assess the extent to which raters' evaluations reflected the variability, in terms of performance, of the data-set we analyzed the variance of the scores for each unitizing technique. A Brown-Forsythe test was run to compare variances among the techniques for both the task and social aspects of communication and for the global scores. For the *Task* dimension, a marginally significant difference was found among the AUT and EST units (p = .057). For the *Social* dimension, no meaningful difference was observed (p = .11). When considering the scores to all questionnaire items, a statistically significant difference was found (p = .04). In all cases, variance was higher for EST units.
Our findings show how both techniques can lead to consistent scores among different raters. This comes as no surprise, as automatic, fixed-window techniques are common-practice approaches [38]. When investigating difference in variances of scores across techniques, a statistically meaningful difference was observed when considering all questionnaire items (i.e. the *global* score), whereas no significant difference was found for the *task* and *social* dimensions separately. It should be noted that, motivated by the goal to compare these results on solo performances with those on social interactions described in [19], the *task* and *social* categorization of items was arbitrary. All items were designed to investigate, in fact, different language functions that could be observed in the performances. For *global* scores, meaningful differences were observed for units coming from the two techniques, suggesting an effect of the technique on the assessment of communicative behavior. The videos we unitized to test our technique depicted 10 different contestants: some did not make it to the second round of the contest, some did and one of them won. As a consequence, different levels of ability and communicative effectiveness were represented in the videos, eventually leading to variance in the scores assigned to videos depicting different contestants. In this sense, the (significantly) higher variance of the EST units illustrates how a cognitive-inspired technique for unitizing affords a more thorough annotation of public speaking performances, in terms of effectively conveying information about a topic (in this work, the *task* dimension) and of entertaining the audience (here, *social*). The results show how an approach for unitizing based on the characteristics of the event to be unitized leads to more thorough evaluations and annotations, therefore overcoming the limitations of a fixed-window approach. In conclusion, the results hereby illustrated pave the way towards an automatic, cognitive-inspired unitizing technique.

2 Testing the approach in a social interaction context

In, [19] we explored the viability of our approach by comparing it with different unitizing techniques, namely: interval coding and continuous coding. With the aim of testing the approach in the domain of social interaction, we selected cohesion as a testbed. The reason behind our choice is that cohesion is a multidimensional construct having five recognized dimensions (task, social, belongingness, group pride, and morale) [86], involving both emotions [69] and goals [61]. As a consequence, cohesion was addressed by studies adopting different unitizing techniques and concerning both psychological and computational aspects, thus affording theoretical comparisons.

Whilst psychologists and sociologists investigated all five dimensions, at the present computer scientists started to address computational methods for automatic analysis of the task and social dimensions (e.g., see [44, 75]). In addition to investigating the effect of unitizing on annotation and exploring the viability of our approach, our work further contributed to this research direction.

In affective computing, emergent states such as cohesion are often investigated by relying on coding schemes. In fact, to capture dynamic group phenomena, research often relies on behavioral coding schemes (see [12] for an overview). These are systems for observing behavior, measuring their occurrence (e.g., number of smiles) or intensity (e.g., warmth) [7]. Prior coding, often the material is unitized, and the unitizing approach may vary depending on the goal of the study, on technical constraints, or on the specific research question at hand [60]. In this study, we investigated whether the coding of social interaction through coding schemes can be influenced by the unitizing approach and whether and EST-based unitizing can lead to more reliable coding and annotation.

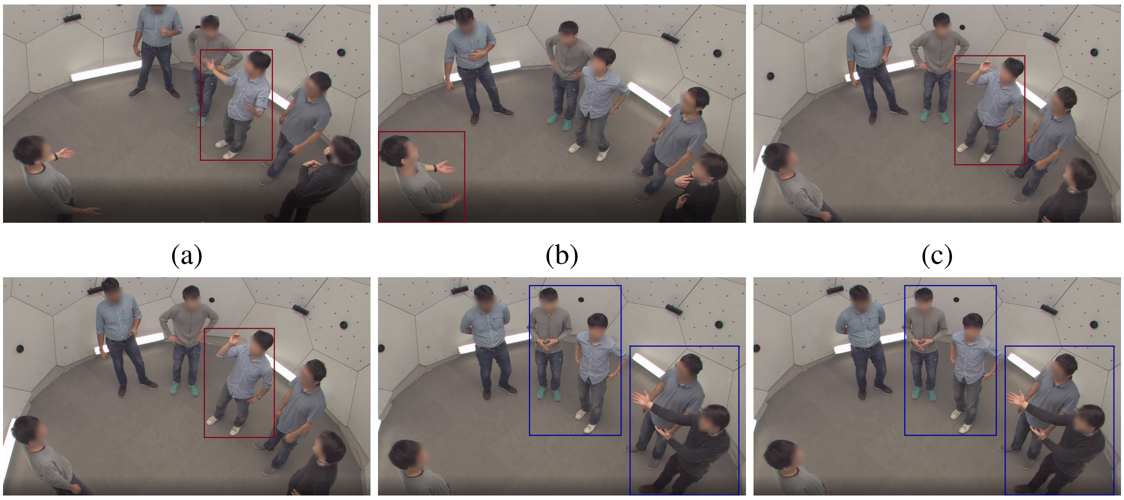
The table 8 illustrates how each change category was operationalized in the specific domain of cohesion annotation. The figure 5 provides further clarification.

Table 5: The left column reports the changes categories as described in [104]. The column in the middle shows how these changes were operationalized. The right column provides an example of occurrence of each change.

Changes in EST	**Changes in our EST-inspired technique**	**Example in group interaction**
C1. Time	Timing and rhythm of the interaction	Group members start gesticulating fast
C2. Space	Motion direction	Group members all move their heads towards the speaker
C3. Objects	Interaction with objects	Participants dismiss a tool they were using
C4. Characters	Character location	One group member leaves
C5. Character interaction	Interaction patterns	Group members start mocking each-other
C6. Causes	Causes and appraisal	Something happens as a consequence of a new state of affairs
C7. Goals	Goals fulfilled, dismissed, or replaced	Group members stop paying attention to the speaker

Our technique derives boundaries from changes in 7 situational dimensions, inspired by the EST but fine-tuned for unitizing group interaction. The most remarkable difference regards the “time” category. In our technique, time is conceptualized as the timing, or rhythm, of the interaction, instead of changes in temporal reference of the scene.

To test our approach in the domain of cohesion, we conducted an online perceptual experiment to compare three unitizing techniques with respect to how they affect:



(a) (b) (c)

(d) (e) (f)

Figure 5:
Example of EST unitizing. In each panel, a bounding box identifies the changes in the scene. In the upper row, only the speaker changes from frame (a) to (b) and from frame (b) to (c). The changes all fall into the same category, i.e., character interaction (red), therefore the frames all belong to the same event. In the lower row, three different categories of changes are observed: from frame (d) to frame (e) the speaker changes (red), the motion direction changes (blue), and the goals of the players change as one of them joking distracts other players from the game. Hence a boundary is placed between frame (d) and (e). Again, only one type of change (movement) is observed between frame (e) and (f), belonging to the same event.

- The average agreement on the raters' scores;
- The extent to which these scores reflect the intrinsic variability of the data-set;
- The loss of information with respect to the scores an expert rater gave to the whole (non-unitized) interactions. That is, a rater scoring a coding unit does it on the basis of information which is limited with respect to a rater who scores the whole interaction [39].

The EST unitizing was expected to provide segments that are perceived as more natural and understandable by annotators and coders. Moreover, we hypothesized an effect of the unitizing technique on the quality of the annotation.
Therefore, a pool of external observers rated task and social dimensions of cohesion on coding units obtained by applying the three different unitizing techniques. The stimuli used in this study are audio-video recordings from the Panoptic data-set [47]. Panoptic is a multimodal publicly available data-set with the following features: natural interactions having rich and subtle non-verbal cues, small groups of up to 8 people, and a large number of camera views (up to 521).

The data-set includes recordings of people playing several social games in small groups. In this study, we focused on the *Ultimatum* game. This game is often used in experimental economics and psychology to study conflict and cooperation. Rules are very simple: one or more players, the *proposers*, are given a sum of money to be split with another player(s), the *responders*. Proposers and responders have a limited amount of time to discuss on how to split the money starting from a proposal done by the proposers. At the end of the established time, if the players agreed on the split, they gain the money, otherwise they loose it. The Panoptic data-set includes 5 Ultimatum's sessions involving each one from 3 to 8 players. A visual inspection of the sessions was performed by an expert psychologist. This resulted in the selection of 12 interactions displaying a variety of behaviors related to the social and task dimensions of cohesion and spanning a broad range of cohesion intensity. Selected interactions involved 3 to 7 players. Interactions were 46 to 57 seconds long (M=53s, SD=3.3s). We discarded interactions involving 8 players because due to the large amount of persons simultaneously acting in the scene, occlusions often occurred making the job of the raters difficult.
To address continuous coding we selected the unitizing technique specified in the ACT4Teams coding scheme (*ACT*). A new unit is created whenever the speaker changes, whenever a speaker utters several statements expressing a thought, whenever the main argument changes, or whenever the speaker talks for longer than 20 seconds. Videos were manually parsed to generate coding units according to this technique. Concerning EST, a trained psychologist watched each interaction and unitized it according to the principles of CEST. Concerning interval coding, we adopted three different sizes for the fixed-window: 8s (*AUT8*), 15s (*AUT15*), and 21s (*AUT21*), respectively. These values were chosen by taking into account previous work on analysis of social interaction in small groups (15s as in [38]), and the average duration of the coding units obtained by applying ACT (8s) and EST (21s). The conceptual model of cohesion by [13], originally applied to sport contexts, is at the basis of most questionnaires used in group cohesion research (e.g., [36, 44, 24, 27, 99]). These questionnaires enable the assessment of cohesion over its dimensions both in first and in third person (i.e., as self-perception and from the point of view of an external rater). We pooled together items from [44, 99, 13] to create a questionnaire containing 10 items organized in two subscales, concerning the task and social dimensions of cohesion. The Likert items of the adopted questionnaires consisted of 7 [44, 99] and 9 [13] points, respectively. In this study, we adopted 5-points Likert items – from 1 (*Not at all*) to 5 (*Yes, definitely*) – as previous studies argued how a 5-points scale can make reading all answers easier for the responders [31]. Table 6 reports the scale used and provides the source for each item. Item from [13] was reported in third person and it was inverted.

Ninety-nine persons (37 males, 60 females, 2 preferred not to specify their gender) voluntarily participated in the study. They were mainly recruited via email advertisements at several universities and research centers. Ratings were collected via a web application in an anonymous form. First, raters were welcome, and given the instructions for their task. Then, they were asked to enter information about age and gender. No information about the Internet connection (e.g., IP address) was tracked nor collected. Participants were informed on data protection policies.

Table 6: The questionnaire on cohesion used in this study.

Task dimension	**Social dimension**
Do you feel that group members share the same purpose, goal, intentions? [44]	Were group members open and frank in expressing ideas/feelings? [99]
Do group members give each other a lot of feedback? [44]	How engaged in the discussion do group members seem? [44]
Do group members seem to have sufficient time to make their contribution? [44]	Do group members appear to be in tune/in sync with each other? [44]
Do group members have conflicting aspirations for the team's performance? [13]	Do group members listen attentively to each other? [44]
Do group members respect individual differences and contributions? [99]	Does the group seem to share responsibility for the task? [44]

Finally, they started to watch the coding units to be rated according to the items in Table 6. Coding units were randomly administered. Next to the video showing the coding unit, a screenshot displayed the group of people to focus on. Raters could leave the experiment when they wished. The total number of gathered ratings was 1771, we discarded 23 ratings due to wrong answers to the honey pot questions appearing in the questionnaire each 10 coding units we asked to annotate. Such questions were used to detected whether the rater was still paying attention to the task. The average number of ratings for each rater was 15. Some raters contacted the experimenter to provide their feedback about the experience and reported having had difficulties to fully understand the following two items of the questionnaire: *Do group members respect individual differences and contributions?*, and *Does the group seem to share responsibility for the task?*. For this reason, these items were removed before running the analysis. Inter-raters reliability was assessed[6] by means of a one-way average-measures (ICC) to evaluate whether raters agreed in their ratings of cohesion across unitizing techniques (AUT8, AUT15, AUT21, ACT, EST). Table 7 shows the obtained values of ICC and the respective 95%-confidence intervals. All the resulting ICCs were in the *excellent* range [56], indicating that raters had a high degree of agreement and that both the task and the social dimensions of cohesion were similarly rated across raters independently by the unitizing technique.

Table 7: ICC values and Confidence Intervals for the task and the social dimension of cohesion

Unitizing	**ICC and 95%-CI**	
	Task	**Social**
AUT8	0.96 (95% CI [0.94, 0.97])	0.94 (95% CI [0.92, 0.96])
AUT15	0.95 (95% CI [0.93, 0.97])	0.97 (95% CI [0.95, 0.98])
AUT21	0.98 (95% CI [0.97, 0.99])	0.96 (95% CI [0.94, 0.98])
ACT	0.96 (95% CI [0.95, 0.97])	0.97 (95% CI [0.95, 0.98])
EST	0.96 (95% CI [0.94, 0.98])	0.97 (95% CI [0.95, 0.98])

The three unitizing techniques (including the three instances of interval coding) can be therefore

[6]The significance threshold for all the tests in this study was set at .05

considered a viable option to achieve suitable ratings for use in hypothesis tests on social and task dimensions of cohesion. For further analysis, the average of the raters' scores is assigned to each coding unit.

To assess the extent to which cohesion scores reflect the variability of the data-set, we analyzed the variance of the scores for each unitizing technique. For both dimensions of cohesion, we first checked for differences between the three instances of interval coding techniques (AUT8, AUT15, and AUT21). Then we compared the interval coding technique which better reflects variability with ACT and EST.

Task: A Brown-Forsythe test detected a significant effect of unitizing technique ($F(2,85.06)=5.43$, $p=.006$). Post hoc comparisons detected a significant difference for AUT21 ($SD=2.49$) and AUT08 ($SD=1.99$), $p=.015$, and for AUT21 and AUT15 ($SD=1.51$), $p=.005$. No significant difference was found between AUT8 and AUT15 ($p=.52$). A Bonferroni correction was applied to account for multiple comparisons.

Social: A Brown-Forsythe test was conducted. A significant effect of unitizing technique was found ($F(2,113.67)=27.20$, $p<.001$). Post hoc comparisons (Bonferroni correction applied) detected a significant difference for AUT21 ($SD=1.61$) and AUT08 ($SD=1.5$), $p<.001$, and for AUT21 and AUT15 ($SD=1.86$), $p<.001$, and for AUT8 and AUT15 ($p<.001$).

Results show that AUT21 reflects variability better than AUT08 and AUT15, having the highest SD for the task dimension and being comparable with AUT15 for the social one. We therefore retained AUT21 for subsequent analysis.

Task: A Brown-Forsythe test was run to compare variances of ACT, AUT21, and EST. A significant effect of unitizing technique was found ($F(2,70.38)=12.48$, $p<.001$). Post-hoc comparisons (Bonferroni correction applied) detected a significant difference for ACT ($SD=1.28$) and AUT21 ($SD=2.49$), $p<.001$, and for ACT and EST ($SD=1.72$), $p=.003$. No significant difference was found between AUT21 and EST ($p=.07$).

Social: A Brown-Forsythe test was run and a significant effect was found ($F(2,104.91)=22.81$, $p<.001$). Post-hoc comparisons (Bonferroni correction applied) detected a significant difference for ACT ($SD=2.15$) and AUT21 ($SD=1.61$), $p<.001$, and for AUT21 and EST ($SD=1.89$), $p<.001$. No significant difference was found between ACT and EST ($p=.055$).

Results show that EST and AUT21 reflect variability in the same way and better than ACT for the task dimension. Concerning the social one, there is no significant difference between EST and ACT, and both outperform AUT21. EST thus overall better reflects variability in both dimensions of cohesion.

To assess the effect of the unitizing technique on loss of information, we compared the scores obtained with each technique with the scores provided by an expert rater who watched the whole non-unitized interaction. For each interaction, Mean Square Error (MSE) was computed between these scores. We carried out this analysis on ACT, EST and the better-performing AUT technique (AUT21). Due to a deviation from a normal distribution of MSE for AUT21 (Shapiro-Wilk test, $W=.80$, $p=.02$), a Kruskal-Wallis test was conducted to examine the differences on MSE according to the unitizing technique. No significant difference ($\chi^2=2.85$, $p=.24$, $df=2$) was found. MSE did not deviate from a normal distribution (Shapiro-Wilk test). A Bartlett test of homogeneity of variances indicated that the assumption of homoscedasticity had been violated (p=.004). A Welch's ANOVA was therefore conducted. A significant effect of unitizing technique was found (F(2.0,14.5)=4.76, p=.03). Post hoc comparisons using the Games-Howell test indicated that MSE for ACT (M=25.41, SD=19.37) was significantly different than AUT21 (M=6.92, SD=5.45), p=.037. EST (M=13.80, SD=13.12) did not significantly differ from ACT and AUT21. Results show that for the social dimension of cohesion, ACT deviates most from the scores given to the non-unitized interaction.

What is more, we decided to investigate in more depth the performances of the unitizing techniques by proceeding coding unit by coding unit. Concretely, we ranked the coding units in order of increasing the standard deviation of the raters' scores. For each unitizing technique, we then computed curve C representing the number of coding units falling below the n-th percentile of standard deviation for n ranging from 5 to 100, step=5. Finally, we compared such a distribution with the ideal distribution where all the coding units generated by a unitizing technique occupy the first positions in the ranking. For performing the comparison, we computed the ratio between the area under curve C and the area under the curve representing the ideal distribution. We obtained the following ratios: for the task dimension, 0.57 for EST and ACT, 0.74 for AUT21; for the social dimension, 0.60 for EST and AUT21, 0.65 for ACT. Concerning agreement, analysis shows that all the unitizing techniques represent a viable option.

Regarding cohesion scores, EST outranked the other techniques in reflecting variability of cohesion in the units for both dimensions. Whereas the task dimension is better assessed when longer units are observed (EST and AUT21 outperformed ACT), our results do not support the same idea for the social dimension, as shorter units (ACT) provided greater variability than longer ones. We think this can be ascribed to raters needing more time to figure out task dynamics than the social one from the players' behaviors. Indeed, raters did not know the rules of the Ultimatum game. For this reason, short coding units could appear not enough "readable" for raters in terms of assessing whether players share goals, intentions, and make contribution. Moreover, following [85], observing an instance of task-related behavior (e.g., turn-taking), requires at least two individual contributions lasting averagely 2s each, whereas emotion recognition can occur in a shorter time (300 ms can be enough) [34]. This confirms that whilst short units were not suitable for the task dimension (especially ACT), they could instead be leveraged for the social one. As a consequence, when trying to evaluate both dimensions of cohesion, a flexible (in terms of time-window) technique is expected to lead to better evaluations. In this study, EST's better performance in reflecting variability could be ascribed to the higher variability in units duration (M=21, SD=10), in line with the idea illustrated in [104] of event perception as a flexible process, that can be fine or coarse grained according to the scope and goals of the perceiver. Concerning the ranking of the coding units, the first 10% of entries (i.e., the first 12 entries) reflects the results discussed above. For the task dimension, 6 units belong to the AUT21 category, 5 belong to EST, and 1 to ACT. For the social dimension, 6 units belong to the AUT21 category, 3 belong to EST, and 3 to ACT. Interestingly, EST units are also diverse in the changes categories they contain with respect to the dimensions of cohesion. The same EST unit had the higher ranking (i.e., lower standard deviation) for both dimensions. This unit contained changes both in character goals and in their interaction. What is more, the next EST entries are not the same entries for the task (4 entries) and social (2 entries) dimensions. With all 4 task entries containing goal changes but only one also containing character interaction change, and both social entries containing changes in character interaction but not in character goals. This diversity in changes distribution aligns with the definition of task and social cohesion provided by [18], with the former indicating group goals and objectives and the latter indicating group members concerns towards relationships within the group. Our results suggest that basing unitizing on the course of the interaction over time (i.e., on changes in the interaction), rather than on time only (AUT techniques) or on behaviors (ACT) can help tackle both the task and social dimensions of group cohesion. An automated EST-based unitizing should focus on such changes primarily.

To conclude, the results of the preliminary studies presented in [19] and [20] highlight a) the effect of the unitizing technique on raters' agreement and scores and b) how the theoretical approach CEST is based on is a feasible possibility to perform unitizing of behavior.

IV Algorithm for semi-automatic multilevel segmentation

Our algorithm is inspired by the EST [108] and, more specifically, by the role of changes in driving boundary perception. The technique is thoroughly outlined in [19] and [21]. This section focuses on its functioning and provides a work-flow describing how to use it for segmentation. Figure 6 outlines the functioning of the algorithm.

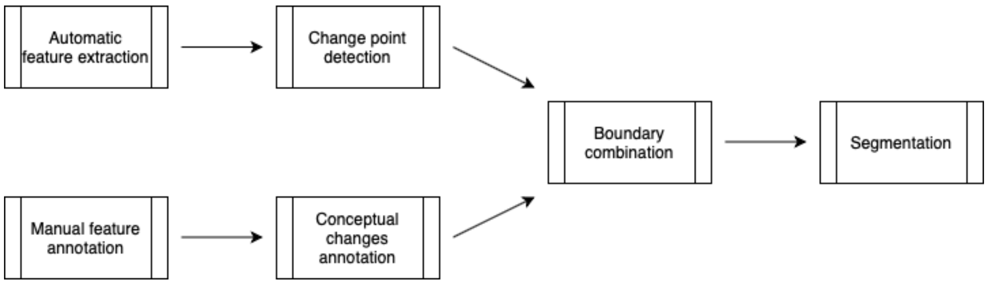


Figure 6:
The functioning of our technique

1 A general purpose multilevel algorithm

The algorithm is, by design, general-purpose. It was created with the aim to provide a tool for researchers and practitioners dealing with the need to segment their movement data. Although the testing described in this manuscript, both in section III and V was carried out on video-recorded and motion capture data, its scope of application is broader, due to its theoretical and technical foundations. Theoretically speaking, the technique is general-purpose as it was inspired by a theory that describes how cognitive segmentation is performed, spontaneously and effortlessly in every-day situations, despite the specific current situation. This process, i.e. the perception of the structure of time [105] is so pervasive in cognitive functioning that even 10 months old infants have been found perceiving boundaries in the current situation [8]. Technically speaking, it is general-purpose as one of the major steps towards segmentation is the identification of change points through change-point detection algorithms. As shown in [6], change-point detection (here, CPD) algorithms proposed in time-series analysis research, are general-purpose in their nature, as change-point detection is a cross-disciplinary field, with applications ranging from medical condition monitoring, climate change, and so on.
Following the EST illustrated in II, the segmentation performed through the technique is based on the perception of changes in the current situation. Seven different change categories can be observed, described by the EST. Each change category corresponds to a specific characteristic of ongoing situation, for instance changes can be detected in the location of characters, in the interaction of one or more characters with objects, in the timing of movements. Furthermore, a distinction can be made between low-level and conceptual changes. In this the technique maps

the approach described in [17], that conceptually divides movement features into low-level, mid-level, and high level features. In this sense, the technique is multilevel. Table 8 maps each change category with its level.

Changes in EST	**Level**
C1. Time	Low-level
C2. Space	Low-level
C3. Objects	Low-level
C4. Characters	Mid-level
C5. Character interaction	Mid-level
C6. Causes	High-level
C7. Goals	High-level

Table 8: The left column reports the changes categories as described in [104]. The right column shows how these changes were clustered into low-level, mid-level and high-level conceptual changes

Changes from C1 to C3 are low-level feature changes, C4 and C5 are mid-level features changes, and C6 and C7 are high-level, conceptual, features changes.
Changes are the starting point towards segmentation: the combination of changes is, in fact, the major novelty of the technique in segmentation research. Inspired by the EST [104], segment boundaries depend on changes and on the fact that boundary perception depends not only from perceiving changes but from perceiving changes from the same category at the same time. More specifically, according to the theory, the chance of a boundary being perceived spikes after detecting changes from three different categories. This combination, however, is not merely a sum of such co-existing changes. As conceptual features have, according to research [63], a higher power than low-level features in driving boundary perception, the sum of co-existing changes is weighted, with conceptual changes scoring double.

2 Workflow

In this subsection, we will provide an outline of the functioning of our technique, also sketched in figure 7. The technique follows four major steps:

- **Feature analysis or annotation:**
  Features are either automatically analyzed or manually annotated. Before this step, each feature of interest needs to be associated with the change category (see II) it pertains to.
- **Change detection:**
  In this step, changes are identified in movement. Changes for each of the 7 EST categories are analyzed separately. For change categories that can be automatically analyzed, change

point detection algorithms are applied to detect meaningful changes in the data stream. For higher-level changes, i.e. causes and goals, this step is strictly linked with the previous one as changes are manually annotated.

- **Boundary combination:**
  Once boundaries have been identified for each change category, they need to be combined in order to perform segmentation. Weights can be adjusted according to the specific research question at hand; however, weighing conceptual features higher than low-level features maps current research on time structure perception [62], [63].

- **Segmentation:**
  When change-points from different categories have been combined and therefore boundaries have been defined, they can be exploited to actually trim the data.

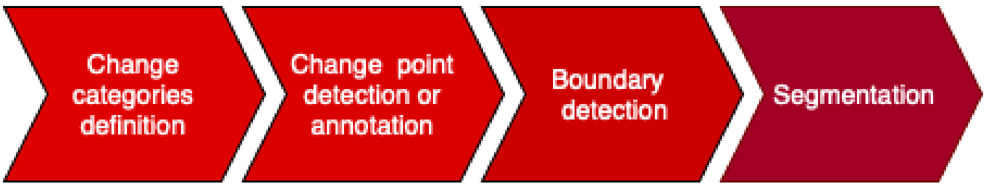


Figure 7: Workflow to achieve segmentation through our technique. Features from each change category are extracted either automatically or manually. After, boundaries are detected through CPD algorithms or via manual annotation. Boundaries are then combined and segmentation is performed when boundaries from different change categories align.

2.1 Feature analysis or annotation

To place boundaries in a data stream, the first step is to select a set of features that a) suitably describe the data and b) are compliant with the change categories posited by the EST. Thus, changes in the data stream are modeled through changes in the features belonging to the different categories.. Among these, some may need to be manually annotated, whereas some may be possibly extracted automatically from a given data-set. For conceptual features, manual annotation may be the best option for two main reasons. The first is that automatic detection of such high-level features, for instance of goals, is still under debate in many research areas, due to the complexity of the topic [52]. Fully automatic approaches are, to the best of our knowledge, not available yet in a form that could be seamlessly integrated with our technique. The second motivation comes from the fact that, notwithstanding the general purpose of the technique, the main area of application we foresee for the output of the work is affective computing. As described in II, in affective computing segmentation is often performed prior annotation or upon the creation of a data-set. In such situations, we see our technique as a tool to lessen the burden of researchers by lifting them a full manual segmentation while also keeping the control of the

parsing of their data-set when it comes to conceptual, high-level features that usually are the core topic of a specific study or data collection. Think, for example, of a study on group cohesion [18]. Group cohesion is a multi-faceted emergent state, that may require a multi-level approach to be efficiently investigated. A researcher working on such a topic might be interested in parsing her data through the combination of low-level, mid-level, and high-level features. However, whereas low-level and mid-level features can be automatically analyzed to extract change points despite the specific field of application, when it comes to high-level features a manual annotation ensures the high-level components of the research topic at hand are thoroughly analyzed in the specific context of application. In the cohesion researcher case, both synchronization [44] and goal alignment [49] can be leveraged to investigate cohesion in a group. While the former can be fruitfully addressed automatically, the latter, however, may require manual annotation.
Upon deciding the features to extract automatically and those to annotate manually, each feature needs to be mapped with one of the 7 different EST change categories. It must be noted that multiple features can be mapped to the same category, for example both turn taking and F-formations [50] can be considered C5 changes, and that one or more change categories can be neglected for a specific case study, for instance if no object is detected in a scene C3 will not be considered for segmentation. Table 9 gives a general description of how EST change categories can be used to cluster different changes in a given scenario.

Changes in EST	**Changes in our EST-inspired technique**
C1. Time	Timing and rhythm of the interaction
C2. Space	Motion direction
C3. Objects	Interaction with objects
C4. Characters	Character location
C5. Character interaction	Interaction patterns
C6. Causes	Causes and appraisal
C7. Goals	Goals fulfilled, dismissed, or replaced

Table 9: The left column reports the changes categories as described in [104]. The right column shows how these changes were operationalized in our EST-based technique

2.2 Change detection

The identification of changes in the data is initially performed independently for each feature and change category. After each feature is mapped with its EST change category (see Table 9), a decision needs to be made on how to identify changes in the scene. As far as conceptual changes are concerned, change detection of conceptual features coincides with the manual annotation of such features. For instance, in the case of goal annotation, the researcher will set a range of possible goals that the character(s) will hold in the scene to be segmented and manually annotate

Goal	Desired state of affairs
understand task	all the players know what they have to do to and the rules of the game
get cue	the players find all the cues that are required or useful to solve the enigma or complete the task
solve task	the players have completed the task, have given the required response or solved the enigma

Table 10: Goal operationalization in the GAME-ON Dataset

them and how they unfold in the scene. In this case, annotating goals coincides with annotating changes in goals. In our belief, goal annotation can be performed by relying on the notion of goals as defined by [95]. Goals in their view are *mental states representing preferred progressions of a particular multi-agent system that the agent has chosen to put effort into bringing about*. Such vision was adopted to frame group goals as its general enough to work across different fields as it refers to a generic definition of *agent system*. For instance, [67] leverage this approach to design agent models for the domain of drama. Take, for instance, the annotation of C7 changes, i.e. goals, in a data-set such as the GAME-ON data-set [70]. The data-set contains motion capture and video recordings of groups of players participating in an Escape game. The game is composed of 5 different tasks that teams have to complete to win the game. For instance, at the beginning of the game participants have to find a box and the keys to open it. The annotation of goals, and hence goal changes can be carried out by formalizing the states of affairs desired by the players. In this, three main goals can be identified in the situation, as illustrated by table 10. The desired state of affairs is therefore for all participants to find the cues and the goal is to get them. For automatically extracted features, change detection is performed by relying on change-point detection algorithms. These algorithms are general-purpose methods, from time-series analysis research, that afford the identification of abrupt changes in a time-series [6]. In our technique, the features from each change category are treated as independent time series, in order to identify change points for each specific category. Chapter V offers a detailed analysis on five different CPD algorithms that can be leveraged to automatically detect changes in a time-series, along with hyperparameter optimization results for each algorithm.

2.3 Boundary combination

According to the perceptual studies carried out by [104], chances of boundaries being perceived spike when changes are observed in three different categories at the same time. As a consequence, our technique places boundaries when, following the two aforementioned steps, changes are detected, simultaneously, across three different categories. It must be noted that boundary perception does not coincide with the perception of changes in three categories, but rather it becomes significantly more likely. As a consequence, the study presented in section V offers data on the optimal change threshold that can be applied to a specific data-set for placing boundaries.

However, both preliminary studies illustrated in section III have adopted a 3 changes threshold. In addition to this, conceptual changes weigh differently, i.e., they are scored doubled meaning if a goal change is detected only one other change needs to be identified at the same time. This follows current literature on segmentation describing the major role of goals in driving boundary perception, as described in II.

To achieve boundary combination, i.e. to transform separate change points into boundaries, the technique finds, for each time-point, detected or annotated change points. When a change point is detected, the time-point is assigned a score equal to 1 for low-level and mid-level changes or 2 for high-level changes. This is performed on the output of change-point detection or of change annotation for each category. As a result, each time-point is assigned a score ranging from 0 to 9, depending on the number of different changes that can be detected simultaneously. It must be noted that a tolerance threshold is applied to change points, so that change points from different categories are considered simultaneous when falling within a given range. In section V results on the best threshold are reported. Algorithms 1 and 2 further illustrate the functioning of CEST: changes from the different change categories are combined and a boundary is placed when the number of changes at the same time point in the data-stream is higher than the selected threshold for combination. The output is an array of boundaries for the data-stream.

Algorithm 1 sum changes

```
for sample in data-stream do
    for every change category do
        for t in time-series do
            if score(t) == 1 then
                if change category is conceptual then
                    score(sample) = score(sample) +2
                end if
                if change category is not conceptual then
                    score(sample) = score(sample) +1
                end if
            end if
        end for
    end for
    return score(sample)
end for
```

Algorithm 2 place boundaries

```
for sample in data-stream do
    if score(sample) > threshold then
        score(sample) is boundary
    end if
end for
return boundaries array
```

2.4 Segmentation

Segmentation is the ultimate goal of the technique. Changes have been combined into overall boundaries that can be leveraged to segment the data. For Mocap data, EyesWeb [16] features a segmentation module. For video, many software options are available, such as ffmpeg[7]. Cognitive segmentation, as described by the EST [108] can have different grains. When segmentation is applied to research, theories can be found that suggest the time-frames to be adopted [2], such as the well-known *thin-slices* approach. However, in line with the multi-purpose nature of our technique, the parameters described in this section as well as in section V can be tuned to adjust window-sizes for one of all the change categories analyzed for a given study.

3 CEST wrap-up

In section II, a framework for segmentation algorithms proposed by [64] was illustrated. Despite being proposed in the field of movement primitives segmentation, the framework can help understand CEST more clearly. Table 11 shows such description. Figure 8 illustrates segmentation through CEST: a set of features describing a movement sequence is analyzed along the different change dimensions postulated by EST [108]. According to the specific data-set, some change categories may be neglected, in this case *objects* and *characters*. In the figure, black lines depict changes and red lines depict segment boundaries. In this case, two boundaries are detected: the first comes from the alignment of two low-level and one mid-level changes, namely *time*, *space* and *location* changes; the second boundary is placed as one low-level, i.e. *space* and one high-level, i.e. *causes* changes are detected at the same time.

Algorithm component	**Definition in CEST**
segment definition	segments are *events* as described by the Event Segmentation Theory [108]
data collection	CEST was tested on dance data-sets recorded through motion-capture technologies, however, the segmentation approach is general-purpose
application specific requirements	being inspired by how cognitive segmentation, a general mechanism, works, CEST is general-purpose
design	preprocessing is recommended and was performed in the evaluation studies (see: section V); the window-size is data-driven, as it depends on the amount and significance of changes in the data
verification	in the evaluation studies (see: section V), F-scores metrics were used to test the performance of CEST

Table 11: Definition of CEST according to the framework by [64]

[7]https://ffmpeg.org

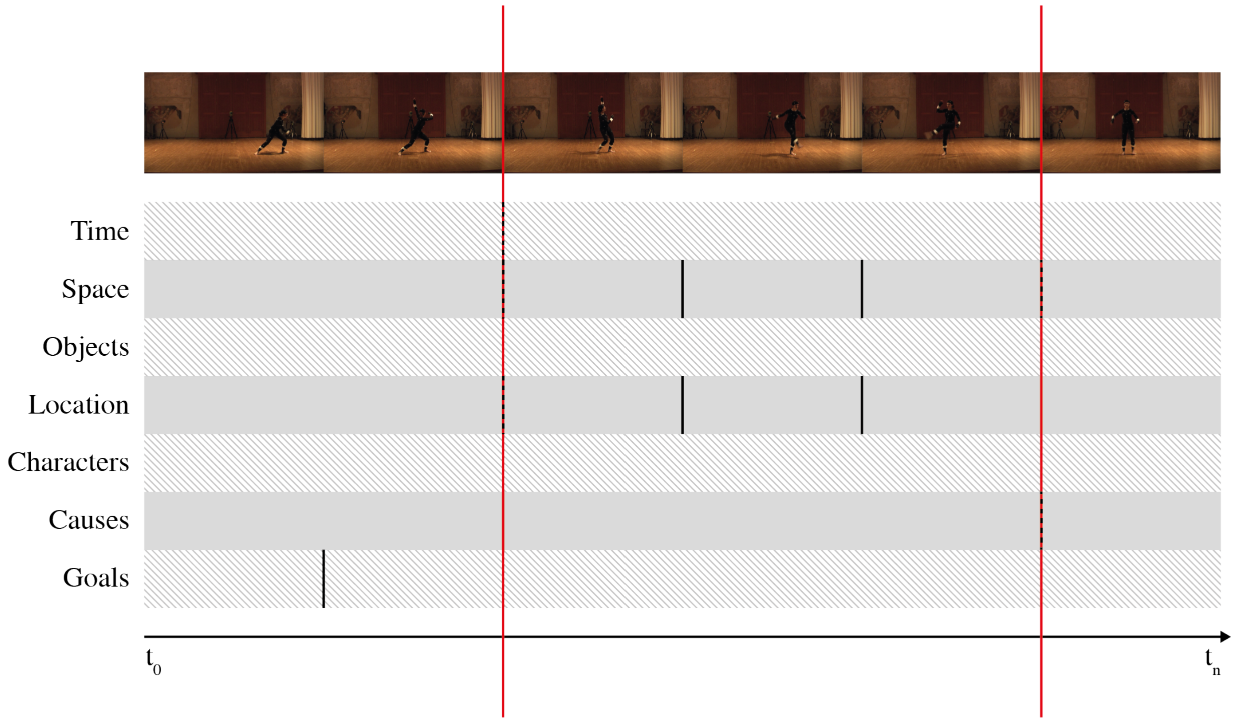


Figure 8:
Segmentation through CEST

V Evaluation of the technique in the context of solo dance performances

This chapter aims at illustrating the experimental evaluation of our technique. Notwithstanding the general purpose of our technique, the initial validation was carried out in the specific context of solo dance performances. This context was chosen as it affords a multilayered analysis of movement features. Dance routines, in fact, tend to be multi-faceted in terms of range of movement and of the affective qualities of the gestures [17]. The decision not to run the technique evaluation on the same single-user data-set analyzed in the preliminary studies, described in chapter III, was motivated by the fact that the solo-stage performances data-set only consisted of video recordings, with speakers often only portrayed in their upper-body. Moreover, no motion capture was available for those performances, thus restraining analysis possibilities for those movements.

Besides evaluating the output of our approach, in our experiment we also investigate the optimal change point detection algorithm for our purpose. The change-point detection step is crucial in our approach, as change detection is the automatic counterpart of change perception in cognitive, everyday-life segmentation. Detecting changes is in fact the starting point towards placing boundaries and, hence, segmentation.

The evaluation phase was designed as follows: after selecting the data-set(s), we identified a set of features that could both be automatically extracted and linked to the EST change categories driving segmentation [108]. Then, one conceptual change category (i.e. causes, indicating changes in the appraisal causes of the current state of affairs, as seen by an external observer), was annotated manually in order to provide change points. Moreover, the automatically extracted features were manually annotated to create ground truth for evaluating the change-points detected through a set of algorithms. Following this, a set of change-point algorithms were selected from time-series analysis research, in order to select the best fit for the purpose of combining multiple change categories in an overall boundary. For each algorithm, change-point detection was performed on both scores and z-scores for each feature selected, thus preventing the results from being influenced by coming from different samples, i.e. two different dancers each performing different sets of routines. Moreover, the output was tested considering different time scales (i.e., different possible segmentation windows for each change point detection algorithm) separately and combined, to explore the role of the granularity of the segmentation on the results. After this, the work focused on boundary combination, with the aim of designing ways to combine detected, or annotated, change-points into overall boundaries, thus mixing the effect of changes across different categories, as according to the EST, happens in cognitive segmentation. At this point, change-points transformed into boundaries were compared with a fully-manual boundary annotation performed by a trained psychology following the EST principles. For each algorithm, precision, recall and F1-scores were measured and compared through Friedman's test [80].

1 Data-sets

As mentioned before, we opted for dance data-sets for the evaluation. More specifically, we selected a data-set of dance movements performed by dancer Cora Gasparotti in a series of studies carried out under the EU-H2020-FET EnTimeMent project [8] and a data-set portraying dancer Marianne Gubri performing contemporary dance routines during a set of studies from the EU-H2020-Wholodance project [9]. Unfortunately, both data-sets aren't currently publicly available. Both data-sets are composed of synchronized video and motion capture recordings of the dancers moving according to different contexts, i.e., expressing different affective states such as anger, impulsivity, and so on. Videos were recorded by means of two professional video-cameras (frontal and lateral view, 1280×720, 50fps) and motion capture data were recorded by means of a 16-cameras Qualysis optical motion capture system ($f_s = 100$Hz)[10]. A total of 20 video and motion captures recordings were analyzed for the evaluation, with a duration ranging from 28 to 169 seconds.

Figure 9:
Dancer Cora Gasparotti (left) and dancer Marianne Gubri (right) on stage at CasaPaganini - InfoMus

2 Feature selection

As illustrated in III, segmenting through our technique requires selecting the features to be analyzed and/or annotated. For this case study, we selected a set of features according to the principles described in IV. The following features were analyzed in their changing patterns:

[8]https://entimement.dibris.unige.it/
[9]http://www.wholodance.eu/
[10]https://www.qualisys.com/

1. *Global Quantity of Movement*: this is the kinetic energy of the movement of the whole body, and it is calculated by taking and analyzing the trajectories of 17 markers on the dancer's body.

2. *Chest Quantity of Movement*: by only calculating the kinetic energy of the movement of one single marker located on a specific joint, i.e. the chest, this feature gives information on the dancer's shifts in the space.

3. *Density of Chest Trajectory*: this is an indicator of whether movement is localized in a small region in the space rather than spanning the whole space, i.e., higher density indicates that the actor has moved in a smaller region.

4. *Directness of Head Movements*: this feature gives information on whether head movement in the space follows straight rather than curvilinear trajectories, i.e., higher directness indicates that the actor has moved along straighter lines.

5. *Causes*: this conceptual feature describes the 3rd person appraisal of the dancer's movements, i.e., the causal attribution an external observer has given to her actions.

Table 12 illustrates how each feature maps with one of the EST change categories. These fea-

Change in EST	**Movement feature**
C1. Time	General Quantity of Movement, Chest Quantity of Movement
C2. Space	Directness of Head Movements
C4. Location	Density of Chest Trajectory
C6. Causes	Appraisal of the movements

Table 12: Mapping between change categories postulated by the EST and movement features

tures were selected as we believe they can provide an effective description of our data-set across different dimensions. Moreover, this feature space allows a representation of the data-set through low-level (C1, C2), mid-level (C4) and high-level (C6) features. As the reader might have noticed, our instantiating of the technique does not take into account 3 EST change categories, namely: objects (C3), character interaction (C5), and goals (C7). For C3 and C5, this is due to the fact that the dancer did not interact with objects nor other characters in the videos. For C7, our choice comes from the idea of only analyzing one conceptual, high-level feature, since such analysis needs to be done manually by an external observer, thus adding arbitrariness to the annotation. As a consequence, causes of movements, rather than the dancer's goals are annotated, as the structure of the videos (i.e., the content of her dance performances) afforded a clear enough understanding of the causes between her movements, such as changes in the context to be represented. Feature analysis for C1, C2, and C4 was carried out through the EyesWeb XMI platform [16], whereas C6 was manually annotated using Elan [103], as illustrated in 3. For automatic feature extraction, 17 out of a total of 61 markers placed on the dancers' bodies were

tracked and analyzed, namely: chest, head, hips, arms, forearms, feet, hands, shins, thighs and shoulders.
Following this, the work proceeded with a manual annotation for the conceptual features and with change point detection for the others.

3 Change annotation

In order to create the ground-truth, videos were manually annotated to detect changes pertaining to those EST categories that could be identified in the scene. This phase had two goals: creating the ground truth to evaluate the performance of the algorithms taken into consideration for our technique and providing boundaries for the conceptual feature C6, i.e. appraisal causes, whose automatic annotation goes beyond the scope of this work.
The videos were annotated by a psychologist[11] that was extensively trained on the EST principles. The annotation was carried out through the ELAN annotation software, allowing multi-layered annotation. ELAN annotation files can also be loaded and manipulated through ANVIL. Figure 10 presents a screenshot from ELAN.

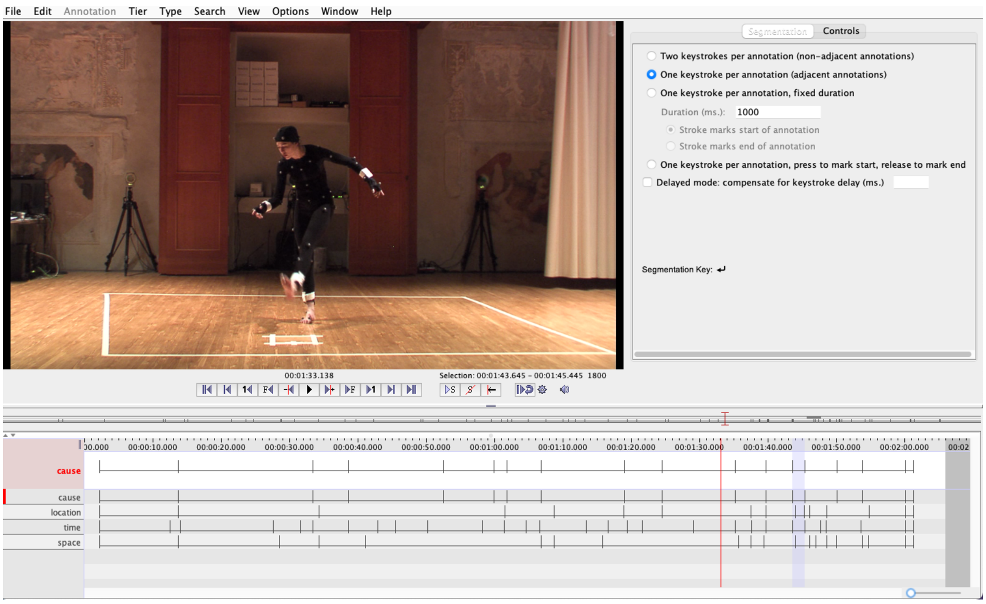


Figure 10: The annotation task.

The videos were annotated as follows:

[11]The psychologist had also previous dance experience. However, her take on the dance movements, as well as notes from the dancers were not taken into account, as our annotation task aimed at collecting an external observer's point of view on the dancers' movements.

- **selecting change categories:** the EST categories that could be identified in the scene were selected to be annotated. For instance, since the character(s) in the scene were never seen interacting with objects, the object interaction category was neglected.

- **annotating changes:** each change category selected was annotated separately leveraging the ELAN segmentation mode. This feature allows placing a change point each time a key is pressed, thus allowing quick and seamless annotation. This phase is detailed later in the text.

- **placing boundaries:** boundaries were placed according to the annotation, where annotating a change from a given category equals placing a boundary in the scene for that category. The output of the ELAN annotation is a text file containing the timing of the annotations in milliseconds formats. In this phase, such timings needed to be transformed into frames to allow comparisons with the CPD algorithms that analyze data at a frame-by-frame level.

- **fusing boundaries:** in this phase, boundaries from the different categories are combined according to the procedure described, for automatic boundaries, in section 5.

As mentioned before, after being selected according to the specific scenarios to be segmented, change categories were annotated one by one, in order to allow comparison with the output of the CPD algorithms for each different dimension. For the single-user scenario, that in this case consisted of a set of dance recordings portraying one performer on stage, the following change categories were analyzed:

- time: a time change is observed when the timing of the action changes, for instance, the agent starts moving slower or faster or the pace visibly changes.

- space: the direction of the movement changes, for instance, the head of the performer points at a different direction.

- location: the position of the performer in the scene shifts, moving from one side of the other.

- causes: the appraisal that can be used to explain the current state of affairs, in this case being the change in the dance performance.

The output of the annotation procedure is illustrated by table 13.

Change category	Number of boundaries detected
C1.Time	85
C2.Space	120
C4.Character location	88
C6.Causes	83

Table 13: The table reports the number of boundaries that were manually detected and annotated. The manual annotation was performed on a total of 20 videos.

4 Change-point detection

4.1 Pre-processing

Before running change-point detection on the data, a pre-processing phase was necessary, due to the specific characteristics of the material at hand. In our case, whereas the movement features were analyzed from the motion capture recordings of the dancers, sampled at 100Hz, the manual annotation, for the cause-related conceptual features and for creating the ground-truth, was performed from a synchronized video, sampled at 50Hz, of the same recording. Therefore, as a preliminary step, motion capture data were down-sampled in order to make the annotation comparable with the output of feature extraction. Moreover, outliers were removed, in order not to have them compromise the analysis. For this task, a Hampel filter, available in the Matlab software, was applied [79] and data were low-pass filtered to remove possible noise. Following the indications in [91], an IIR low-pass filter having a 15Hz cutoff frequency and the transfer function reported in [91] was applied.

4.2 Candidate algorithms

As described in chapter IV, once features are extracted or annotated, change point detection needs to be performed on the data to identify meaningful changes and to place boundaries.
This section will illustrate a study on a set of algorithms from time-series analysis research that was carried out to identify the better option for the change-point detection step that is, indeed, the core of segmentation. Many CPD algorithms are available in the literature to analyze changing patterns in time series (see 3). For our technique, several different options were explored. The rationale behind our selection is that the candidate algorithms need to be widely acknowledged in CPD research and need to have a theoretical approach to change detection that can be mapped with those of the EST. Table 14 gives a snapshot of the algorithms that were taken into consideration.
More specifically, we tested:

Name	References	Core characteristics	mapping with EST
RuLsif	[66]	non-parametric divergence estimation of point density	importance of changes
E-Divisive	[71]	change point detection as clustering	boundaries as the most informative part of a sequence
BOCP	[1]	change point as probability estimation	role of prediction
PELT	[53]	prioritizing minimizing cost function	segmentation as a real-time cognitive process

Table 14: The table demonstrates the candidate algorithms for our segmentation task.

- **RuLsif** [66] is a change-point method based on density estimation. The authors apply a density-ratio estimation method called the unconstrained least-squares importance fitting (uLSIF) developed by [48] to CPD. RuLsif is in fact an extension of uLSIF, which uses relative rather than absolute density ratios, thus obtaining better estimates with less computational costs. The core of the technique lies in estimating the (non-parametric) divergence between the probability density of time-series samples from subsequent segments. To achieve such divergence estimation, the technique uses relative Pearson divergence. The rationale behind considering this technique for our purposes is the importance that divergence of subsequent segments have in boundary detection, which is in line with the idea that cognitive segmentation is, as a matter of fact, change detection [108]. In testing the performance of RuLsif for our purposes, we explored different possible parameters configurations. Table 15 summarizes the parameters we tested for this algorithm.

Parameter	function	range	test
alphas	parameter used for the computation of the Pearson divergence	0 - 1	0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9
window size	temporal scale	10 - 250 frames	10, 25, 50, 100, 150, 250
threshold	the threshold to be applied for detecting a change. This corresponds to quantiles of the divergence score	0 - 1	0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9

Table 15: The table illustrates the parameters that were tested for the RuLsif [66] algorithm. Default values were used for folds and sample size of the segmentation window

- **E-Divisive** [46] detects change points by comparing and measuring the characteristic functions of the distributions of subsequent segments of a time-series. In brief, it finds multiple change-points by iteratively applying the procedure for a single change point. Then, the statistical significance of each estimated change point is measured through a permutation test. The core concept is that characteristic functions uniquely describe a probability distribution and therefore changes in such functions equal changes in the distribution. To summarize, E-divisive segments through a series of steps: a) segmenting the time series into two segments for which the characteristic functions maximally differ; b) determining the significance of the change point by running a permutation test; c) dividing the sequence into separate units according to the detected change points and searching for further change points in each of them. In this, E-divisive maps with the idea that the most informative portion of a scene lies between boundaries rather than within boundaries [104].

Parameter	function	range	Test
minimum inter-onset interval	minimum interval between two different boundaries	10 - 250	10, 25, 50, 100, 150, 200
statistical significance	statistical significance of the difference between samples from different segments	0 - 1	0.001, 0.005, 0.01, 0.02, 0.03, 0.04, 0.05

Table 16: The table illustrates the parameters that were tested for the E-divisive algorithm [71] algorithm.

- **BOCP** [1] is a Bayesian change-point detection algorithm for online inference. It is based on the assumption that a sequence of observations can be divided into non-overlapping units, thus placing change-points into the sequence. The core of the algorithm lies in the estimation of the posterior probability of the current run length for a unit at a given time. What is more, we tested two different instances of OCP, namely the best Bayesian and standard Bayesian. This algorithm was selected for this study as, in [14], it was reported as having the best performance when compared with other common approaches more importantly, as it takes into account the role of prediction in boundary perception postulated by the EST [106].

Parameter	function	range	Test
minimum inter-onset interval	minimum interval between two different boundaries	10 - 250	10, 25, 50, 100, 150, 200
statistical significance	statistical significance of the difference between samples from different segments	0 - 1	0.001, 0.005, 0.01, 0.02, 0.03, 0.04, 0.05

Table 17: The table illustrates the parameters that were tested for the OCP algorithm [1].

- **PELT** [53], or pruned exact linear time, applies optimization techniques to find change point by minimizing a cost function. The algorithm deals with the problem of the increase in the number of change points over time by adding a term to the cost function so that the optimal number and location of change points is retrieved. PELT has a computational cost that is linear in the number of observations. Along with being widely used in change point detection research [40], this algorithm was selected as it works almost in real-time, as in fact does cognitive segmentation, a spontaneous process of the human mind [105].

Parameter	function	range	Test
minimum inter-onset interval	minimum interval between two different boundaries	10 - 250	10, 25, 50, 100, 150, 200
statistical significance	statistical significance of the difference between samples from different segments	0 - 1	0.001, 0.005, 0.01, 0.02, 0.03, 0.04, 0.05

Table 18: The table illustrates the parameters that were tested for the PELT [53] algorithm.

To sum up, we explored five different techniques, for 4 different features, combined into 3 EST change categories, over two different kinds of data and for all we explored two different temporal configurations. According to research on event structure perception [76], [105], cognitive segmentation varies in the temporal span, ranging from fine-grained to coarse-grained segmentation. To map this variability, we decided to investigate the optimal segmentation window for our time structure analysis. With this goal, for each of the 4 different features, we tested each technique over different temporal scales, ranging from 10 to 250 frames.
For each algorithm, we obtained a perscale segmentation and a multiscale segmentation. While the former provides a specific temporal scale, by finding the optimal temporal scale for each feature, the latter gives us information on the saliency of a boundary across different temporal scales. The multiscale analysis, in fact, combines the results of all the different perscale segmentations. For instance, a boundary can be placed at a given time-point when the segmentation window is shorter but can "disappear" as the window is enlarged, or vice-versa. As a consequence, the detection of a boundary, over different time-scales, at the same point can be considered as an indicator of the saliency of that particular instant, whose change matters across different observation time-spans. Multi-scale segmentation, can be argued, provides pivotal change points in a given time-series more accurately than per-scale segmentation. In addition to the CPD algorithms, a fixed-window approach was adopted, here referred to as AUT, as often automatic segmentation coincides with selecting a time-window (see section II). AUT sets a change point every fixed number of samples according to the selected temporal scale (i.e., every 10, 25, 50 samples, and so on). Segmentation was performed on each change category to be compared with automatically extracted changes and with combined changes, i.e. boundaries.
The following section details boundary combination.

5 From change points to boundaries

After change-point detection is performed on each different component, i.e. change category, change-points need to be turned into boundaries, combining changes from different categories. According to the EST [108], boundary perception not only centers around perceiving changes from different change categories, but it also requires such different changes to be co-existent at the same time. A one-minute clip where the only changes lie in the velocity of hand movements might be perceived, by an external observer, as having no boundary as well as one where changes in, for instance, the velocity of hand movements and the location of the character are placed very far in time one from another [104]. As a consequence, our technique performs an *ad-hoc* combination of the outputs of change-point detection, which is grounded on most recent research on cognitive segmentation, stating how conceptual changes (i.e., causes and goals) are more powerful in driving segmentation that low-level changes [63]. Therefore, we propose a combination that weights detected, or annotated, changes in each category, assigning a score of 1 to low-level changes and of 2 to conceptual changes.

EST change	**EST weight**	**weight in our experiment**
C1.Time	1	1
C2.Space	1	1
C3.Objects	1	not considered in our experiment
C4.Characters location	1	1
C5.Character interaction	1	not considered in our experiment
C6.Causes	2	2
C7.Goals	2	not considered in our experiment

Table 19: The table reports the proposed weights to be assigned to boundaries from each change category according to the theory and weights assigned in our experimental study

Before combining change-points, detected changes in the same time-series that were closer than 10 frames were fused, with the boundary being placed between the two frames. For instance, a change-point detected at frame 380 and another one at frame 390 will result in a boundary being placed at frame 385. This post-processing is necessary, to clean the output of CPD from boundaries that are too close, in time, to be both meaningful when compared to boundaries perceived by a human observer [54].
After this, boundaries from different categories were fused. To do so, each frame containing a change point was assigned a score, according to the scoring system described in table 19. For each time-point, therefore, an overall boundary was obtained by combining the scores from the different categories. With this procedure, 5 different possible segmentations were generated from the score of each time point. Scores ranged from 1 when only one low or mid level change was detected to 6, when changes were detected simultaneously in all change categories. According to [104], boundaries are more likely to be placed in a scene when changes are detected across

three different categories, and according to [63], conceptual changes drive boundary perception. However, in this study, all possible boundary combinations were explored, in terms of their ability to replicate manually annotated ground-truth data, as measured by F1-statistics.

6 Results

To evaluate our approach to semi-automatic segmentation, we ran a set of experiments. Statistic analysis on our data had different aims:

- testing the performances of CPD algorithms on our data-set, by measuring F1-scores for each CPD technique on each feature, for scores and z-score;
- measuring the performance of CEST, in terms of F1-scores, by testing it against the ground-truth obtained through manual annotation
- comparing the performance of CEST with that of automatic fixed-window (here, AUT) segmentation
- identifying the best CPD technique for CEST
- comparing F1-scores for perscale and multiscale segmentation
- comparing F1-scores for segmentation based on low-level and mid-level features only with segmentation including conceptual features
- identifying the best parameters for the technique

In our experiment, automatically detected change points and boundaries were compared with the manual ground-truth obtained through the procedure described in section 3. The testing on CPD algorithms was aimed at exploring whether our results confirm previous research on CPD [14]. The comparison between perscale and multiscale segmentation had the goal to assess the role of the temporal scale in segmentation, as research, both in cognitive science [104] and affective computing has stressed the importance of this dimension for behavior analysis. In terms of performance, the data hereby presented are the first attempt at evaluating CEST, although the theoretical approach was proven effective in [19] and [20]. Moreover, we also compared the performance of CEST with that of automatic, fixed-window segmentation, here referred to as ”AUT” as it is a very common research approach to the unitizing problem (see section 2).
Although all the CPD techniques were selected according to their performance in previous studies and to their compliance with EST (see: table 14), the testing was also aimed at providing results on the best algorithm for segmenting through CEST and on the best parameters. All analysis were performed on the two data-sets jointly. Moreover, we combined the data-sets into a test data-set and validation data-set. We explored 100 different test-validation possible combinations,

obtained by combining all recordings whose GT annotation contained a percentage between 70 and 80 of total changes from all the recordings in the data-set.
All the analysis in this section were carried out using JASP statistics software [68]. All analysis were run on scores and z-scores. Here, we report those performed on scores. Results for z-scores are reported in the appendix, as long as tables detailing all the analysis. Here, we summarize the main results.

6.1 Results on changes

Before evaluating the performance of CEST, we tested the performance of each CPD algorithm on our data-sets. Automatically detected changes were compared against ground truth data obtained through manual annotation by a trained psychologist. Outputs of CPD were compared by measuring, for each change category, precision, recall and F1 scores of each algorithm. The

Algorithms	Change categories	Temporal grain	Measures	Scores
• AUT • RuLIF • EDivisive • OCP • OCP best • PELT • PELTP	• C1. time • C2. space • C4. location	• per-scale • multi-scale	• precision • recall • F1-scores	• scores • z-scores

Table 20: The table sums up the testing performed on time, space and location changes.

results are further detailed in the Appendix.
F1-scores were measured, for each component and for each temporal scale by comparing automatically detected changes with changes manually annotated by a a trained psychologist. Tables 21, 22 and 23 show metrics for each change category on validation sets. Results for test sets and for z-scores are reported in the Appendix.

6.2 Results on boundaries

The boundaries obtained through the weighed combination of automatically detected and annotated changes were compared with ground-truth data obtained via manual annotation from a trained psychologist. Table 24 reports the tests performed on boundaries. For all validation and

			AUT	RuLIF	CPD	OCP	OCPB	PELT	PELTP
Temporal scale	perscale	Min. value	0.238	0.223	0.260	0.230	0.241	0.259	0.248
		Max. value	0.294	0.299	0.326	0.354	0.359	0.327	0.308
		Mean	0.270	0.260	0.294	0.309	0.310	0.292	0.278
		standard deviation	0.011	0.014	0.014	0.023	0.022	0.015	0.015
	multiscale	Min. value	0.285	0.318	0.322	0.285	0.316	0.324	0.324
		Max. value	0.402	0.415	0.419	0.369	0.407	0.425	0.417
		Mean	0.338	0.363	0.364	0.324	0.360	0.369	0.363
		standard deviation	0.025	0.021	0.021	0.017	0.018	0.021	0.019

Table 21: Metrics for F1-scores for time change category

			AUT	RuLIF	CPD	OCP	OCPB	PELT	PELTP
Temporal scale	perscale	Min. value	0.206	0.268	0.232	0.174	0.148	0.241	0.228
		Max. value	0.290	0.354	0.322	0.258	0.242	0.341	0.314
		Mean	0.249	0.314	0.279	0.207	0.198	0.293	0.272
		standard deviation	0.019	0.019	0.020	0.014	0.018	0.021	0.020
	multiscale	Min. value	0.285	0.318	0.322	0.285	0.316	0.324	0.324
		Max. value	0.402	0.415	0.419	0.369	0.407	0.425	0.417
		Mean	0.338	0.363	0.364	0.324	0.360	0.369	0.363
		standard deviation	0.025	0.021	0.021	0.017	0.018	0.021	0.019

Table 22: Metrics for F1-scores for space change category

			AUT	RuLIF	CPD	OCP	OCPB	PELT	PELTP
Temporal scale	perscale	Min. value	0.206	0.268	0.232	0.174	0.148	0.241	0.228
		Max. value	0.290	0.354	0.322	0.258	0.242	0.341	0.314
		Mean	0.249	0.314	0.279	0.207	0.198	0.293	0.272
		standard deviation	0.019	0.019	0.020	0.014	0.018	0.021	0.020
	multiscale	Min. value	0.178	0.233	0.208	0.174	0.166	0.212	0.169
		Max. value	0.275	0.301	0.294	0.258	0.240	0.312	0.265
		Mean	0.230	0.268	0.250	0.207	0.201	0.269	0.209
		standard deviation	0.018	0.015	0.018	0.014	0.015	0.021	0.018

Table 23: Metrics for F1-scores for location change category

test sets, F1 scores from all techniques were compared by relying on Friedman's statistics, following [88] and [32]. Analysis were run on results on perscale and multiscale segmentation for segmentations, i.e. boundary combinations, obtained on low-level automatically extracted features only and for segmentations obtained on conceptual level and low and mid-level changes. For each temporal grain - boundary combination, we explored, for each set, the effect of the algorithm on the F1-scores, to evaluate whether using a specific algorithm would lead to a better performance of the technique for segmentation.

Validation sets:

Algorithms	Boundary combinations	Temporal scale	Measures	Scores
• AUT • RuLIF • EDivisive • OCP • OCP best • PELT • PELTP	• with cause changes • without cause changes	• perscale • multiscale	• precision • recall • F1-scores	• scores • z-scores

Table 24: The table sums up the testing performed on boundaries obtained by combining low-level and mid-level features only (without cause changes) and by combining all features.

When analysis were performed on without-causes perscale segmentation, a significant main effect of the algorithm on F1-scores was found for tests run on scores $\chi^2(6) = 396.112, p. < 0.001$ and z-scores $\chi^2(6) = 398.879, p. < 0.001$. Despite F1-scores for without-causes segmentation being lower than with-causes segmentation, the statistics on with-causes segmentation led to similar results, for both scores $\chi^2(6) = 541.208, p. < 0.001$ and z-scores $\chi^2(6) = 537.691, p. < 0.001$.
Moving to multi-scale segmentation, a significant main effect of the algorithm on F1-scores was found for tests on without-causes segmentation run on scores $\chi^2(6) = 444.731, p. < 0.001$ and z-scores $\chi^2(6) = 411.349, p. < 0.001$, and for tests on with-causes segmentation run on scores $\chi^2(6) = 522.524, p. < 0.001$ and z-scores $\chi^2(6) = 408.675, p. < 0.001$. For all possible temporal scale vs segmentation configurations, Conover's post hoc comprarison revealed that F1-scores for AUT were lower than for the other algorithms. This was true for scores and z-scores.

Test sets:
Analysis performed on without-causes perscale segmentation have highlighted a statistically meaningful effect of the algorithm on F1-scores, for tests on scores $\chi^2(6) = 142.853, p. < 0.001$ and z-scores $\chi^2(6) = 135.297, p. < 0.001$ as well as for tests performed on with-causes perscale segmentation scores $\chi^2(6) = 807.615, p. < 0.001$ and z-scores $\chi^2(6) = 689.075, p. < 0.001$. Regarding multi-scale segmentation, the effect of the algorithm was confirmed for without-causes segmentation on scores $\chi^2(6) = 129.005, p. < 0.001$ and z-scores $\chi^2(6) = 110.153, p. < 0.001$ as well as for with-causes segmentation for scores $\chi^2(6) = 421.576, p. < 0.001$ and z-scores $\chi^2(6) = 398.998, p. < 0.001$. For all possible temporal scale vs segmentation configurations, Conover's post-hoc comparison highlighted that F1-scores for AUT were lower than for the other algorithms. This was true for scores and z-scores.

The full results from Conover's post-hoc testing are reported in the Appendix. In all analyses, AUT F1-scores were significantly lower than scores obtained by segmenting through the CPD algorithms.

F1-scores for all algorithms are reported in the appendix, for the sake of brevity, tables in this section only depict the highest-scoring techniques in validation and test sets along with the F1-scores obtained by AUT segmentation.
Overall, OCPB and OCP gave the best performance. Tables 25, 26 and 65 summarize the results of the best performing techniques.

		temporal	**scale**
		perscale	**multiscale**
conceptual	yes	OCPB(OCPB)	OCP(OCPB)
features	no	OCP(OCP)	OCP(OCP)

Table 25: Best performing technique for each configuration, assessed on scores, in the validation sets (test sets in brackets).

algorithm	**temporal scale**	**segmentation**	**F1-scores**
OCPB	perscale	with conceptual features	0.775
OCP		without conceptual features	0.187
OCP	multiscale	with conceptual features	0.779
OCP		without conceptual features	0.187
AUT	perscale	with conceptual features	0.374
		without conceptual features	0.108
	multiscale	with conceptual features	0.290
		without conceptual features	0.101

Table 26: Best F1-scores and AUT F1-scores for the validation set

algorithm	**temporal scale**	**segmentation**	**F1-scores**
OCPB	perscale	with conceptual features	0.801
OCP		without conceptual features	0.190
OCPB	multiscale	with conceptual features	0.789
OCP		without conceptual features	0.188
AUT	perscale	with conceptual features	0.387
		without conceptual features	0.121
	multiscale	with conceptual features	0.271
		without conceptual features	0.116

Table 27: Best F1-scores and AUT F1-scores for the test set

Once the best performing techniques were identified, tests were run to assess whether the temporal grain of the segmentation led to significantly different scores. As table 28 demonstrates, t-statistics ($\alpha = .01$) found no statistically meaningful difference between perscale and multiscale segmentation. Results were similar for all other CPD techniques.

segmentation	perscale	multiscale	t	df	p
with conceptual features	OCP	OCP	-1.472	99	0.072
	OCPB	OCPB	-1.761	99	0.041
	AUT	AUT	8.185	99	1.000
without conceptual features	OCP	OCP	0.808	99	0.789
	OCPB	OCPB	12.004	99	1.000
	AUT	AUT	13.392	99	1.000

Table 28: Comparing perscale and multiscale segmentation

In addition to this, scores were compared in order to assess whether segmentations including conceptual features led to meaningfully higher scores than segmentations based on low-level and mid-level features only. The results, illustrated in table 29 and table 30, indicate that segmentations including conceptual features lead to higher F1-scores.

Without conceptual features	With conceptual features	t	df	p
OCP perscale	OCP perscale	-120.543	99	$< .001$
OCP multiscale	OCP multiscale	-123.276	99	$< .001$
OCPB perscale	OCPB perscale	-182.081	99	$< .001$
OCPB multiscale	OCPB multiscale	-214.630	99	$< .001$

Table 29: Comparing segmentations without and without conceptual features, validation sets

Without conceptual features	With conceptual features	t	df	p
OCP perscale	OCP perscale	-74.051	99	$< .001$
OCP multiscale	OCP multiscale	-72.426	99	$< .001$
OCPB perscale	OCPB perscale	-78.570	99	$< .001$
OCPB multiscale	OCPB multiscale	-114.245	98	$< .001$

Table 30: Comparing segmentations without and without conceptual features, test sets

Tables 31 and 32 show the results of GRID search on scores for the best performing techniques. Results on z-scores are reported in the Appendix. Although CEST is, by design, general purpose, we believe further research is needed to identify the best parameters to leverage CEST for segmentation across different research scenarios, such as group or dyadic interactions. We hope future developments will shed more light on this topic. As far as boundary combination is concerned, for both OCP and OCPB, the mean optimal threshold for all sets was 3 for segmentations without conceptual features and 2 for segmentations with conceptual features. In this case, we hypothesize this threshold will hold across different research aims, as it maps the spike in the probability to detect boundaries after changes in three different categories posited by the EST [104], along with the importance of goals in driving boundary perception [62].

component	temporal scale	threshold	sig.	λ	α	β	κ	window size
time	perscale		0,4	200	0,01	100	1	250
	multiscale	0,1	0,4	200	0,01	100	1	
space	perscale		0,1	100	1	0,01	0,01	250
	multiscale	0,1	0,1	100	1	0,01	0,01	
location	perscale		0,1	50	100	0,01	1	250
	multiscale	0,1	0,1	50	100	0,01	1	

Table 31: Best parameters for OCP, grid search on scores

component	temporal scale	threshold	sig.	λ	α	β	κ	window size
time	perscale		0,1	100	0,01	100	100	50
	multiscale	0,8	0,1	200	0,01	100	100	
space	perscale		0,1	50	1	0,01	0,01	25
	multiscale	0,2	0,1	200	0,01	0,01	1	
location	perscale		0,1	100	100	0,01	1	10
	multiscale	0,1	0,1	50	100	0,01	1	

Table 32: Best parameters for OCPB, grid search on scores

6.3 Results wrap-up

Table 33 summarizes the output of our experiments. Bold letters indicate best options for algorithms, temporal scale and boundary combination, according to our studies on dance improvisation data-sets.

Algorithm	Temporal scale	Boundary combination
AUT		
EDivisive	**with conceptual features**	**perscale**
RuLif		
OCP		
OCPB		
PELT	without conceptual features	**multiscale**
PELT-P		

Table 33: Main results

7 Discussion

To avoid the effect of individual differences in dancers' movements on the evaluation of CEST, the performance of the technique was tested on both data-sets jointly. Firstly, each change point detection technique was tested separately, in terms of precision, recall and F1-scores, against the ground-truth collected through manual annotation. In both validation and test sets, F1-scores were lower than 0.5 for all algorithms. This was true for scores and z-scores. Boundary combination is the core of CEST. Changes to be combined were automatically detected through 7 different change point detection techniques from time-series analysis research. Differently from changes, on boundaries all techniques led to mean F1-scores higher than 0.5, for validation sets and test sets, with highest F1-score reaching 0.8. However, F1-scores were only high for boundary combinations including low-level, mid-level and high-level features, whereas F1-scores for low and mid level features only were lower, with highest mean score being 0.190. One possible interpretation could be that including one manually annotated change led to a rise in F1-scores when comparing boundaries with ground truth data, also manually annotated. Nonetheless, this should have resulted in F1-scores rising as well. Instead, F1-scores for AUT when segmentations on boundaries including conceptual features don't have a statistically meaningful difference from F1-scores on those based on low and mid level features only. This supports the idea of CEST as a multilevel technique, grounding the segmentation on the interplay between different features, illustrated in section IV. This is a novel approach in segmentation research, where segmentations usually leverages one dimension specifically [64]. The scores are promising. The overall good performance obtained by multilevel segmentation suggests the possibility to adopt this approach to different research topics, as it can take features from different levels of analysis into account.
Regardless of the change-point detection algorithm adopted, CEST achieves better performance than the fixed-window AUT approach. This finding is line with cognitive science research challenging the idea of conscious perception as a succession of discrete temporal frames [101] while also confirming the results obtained in the feasibility studies in single [20] and group [19] scenarios. Furthermore, this is in line with research on unitizing, demonstrating the risks behind adopting a fixed-length window to analyze behavior.
The analysis also had the goal to identify the best performing algorithm for CEST. Overall, across validation and test sets, OCP and OCPB [1] obtained the highest F1-scores. It must be noted that both are two versions of the same approach, namely BOCP. The core of this algorithm is prediction: the length of each unit is inferred. This mirrors and further supports most recent research on cognitive segmentation, stressing the idea of segmentation as prediction [83]. Segmenting is comparing one's own predictions with the current situation, always being one step ahead. OCP and OCPB adopt a Bayesian approach, the same probabilistic approach to uncertainty that, according to a very extensive line of psychological research, the brain adopts to manage uncertainty and, ultimately, for prediction [37], [55]. What is more, BOCP is an online technique. Although testing an online version of CEST goes beyond the scope of this work, these results suggest this possibility.
Mean F1 scores from OCP and OCPB obtained through perscale segmentation where then com-

pared with those obtained through multiscale segmentation. No statistically meaningful mean differences in F1-scores were found. The experiments also highlighted that, for the selected algorithms, the optimal threshold combinations were 3 boundaries for segmentations without conceptual features and 2 boundaries when conceptual features are taken into account. This results is in line with the the number of boundaries that, according to the EST, leads to a spike in the probability of event segmentation [104]. F1 mean scores for test and validation sets suggested that CEST performance was better when boundary combination took conceptual features into account, despite the change point detection algorithm. The analysis on the best performing algorithms confirmed this results. A statistically significant difference was detected between low and mid-level segmentation and conceptual segmentation. This further supports the approach of CEST and matches literature on cognitive segmentation positing the importance of conceptual features [63]. Parents know how very young children take pride in dropping objects; dropping an object that is then picked up only to be dropped again lets children repeat the fulfilling experience of achieving a goal. This happens as, from a very early stage, goals are the building blocks of our perception of the continuous stream activity [8]. Conceptual features have a primary role in boundary perception: our results further support this claim.

VI Conclusions

This manuscript has described the attempt to solve a multidisciplinary problem through multidisciplinary research. The output of the work is CEST, a Cognitive Event based Semi-automatic Technique for behavior segmentation. CEST aims at proposing a general purpose technique for all researchers needing to segment their behavioral data prior analysis. However, it was designed after exploring the issue in the specific context of affective computing, as it is a good example of a research field where complex, multi-level and multidisciplinary phenomena are tackled. CEST aims at lifting affective computing researchers from the task of manually annotating their data by offering a tool to perform the annotation of low and mid level features automatically. Our preliminary studies have demonstrated how a fully automatic segmentation, for instance through fixed-length windows, affects the quality of the annotations and, ultimately, the research results. CEST is a technique that, partially, lifts from manual unitizing while keeping the results safe from the effects of automatic segmentation. What is more, our results further support the Event Segmentation Theory, as they show the feasibility of implementing its principles into computational models. This possibility paves the way towards the design of EST-applications for the diagnosis of memory-loss, as event structure perception deficits have been found robust indicators of age-related memory impairments [93] [110].

The design of the technique started from a cognitive theory on how segmentation works in the human mind, namely the Event Segmentation Theory [108]. To perform segmentation, CEST leverages research from the field of time-series analysis: automatic segmentation is performed by relying on change-point detection algorithms. The combination of such changes is achieved by following the EST. The approach of CEST was proved feasible in a set of preliminary studies [19] [20].This manuscript has focused on validating CEST against a data-set of dance movements. Several change point algorithms to be used for this purpose were explored; although all have shown a good performance when compared with ground-truth data, the best performing were two algorithms adopting an online Bayesian approach to change point detection, namely OCP and OCPB. Interestingly, the approach matches cognitive research on the perception of the temporal structure of events. The brain, according to such theories, is a Bayesian predictive machine [33]. The work also provides the best parameters for the technique. From the analysis, when each of the different change components processed for boundary detection are segmented according to their optimal segmentation window (what we referred to as *perscale segmentation*), different components require different time scales to be optimally segmented. This further supports the idea of using CEST, a technique which is multilevel, in the sense that it combines different features and different levels of analysis to achieve segmentation. Also, these results offer future research directions as they demonstrate how the selection of the temporal granularity of analysis cannot be solved as a ”one size fits all” problem.

Rather than proposing the ultimate segmentation technique, the research on CEST has highlighted some of the gears that such technique would require. CEST is, at least currently, not fully automatic. Robust methods to automatically analyze conceptual features from movement may help CEST offer this option to researchers hoping for a fully automatic cognitive-based technique. CEST proved itself successful on our solo dancers data-sets. To further explore semi-automatic cognitive segmentation, future research is needed to test CEST on a broader range of behavior data-set, for instance on social interaction data. Pivotal research question arise from this possibility, such as how single-users features analyzed for change point detection can be combined to achieve an overall event segmentation of a social interaction where many characters are simultaneously involved. To reach this goal, we advocate for a multidisciplinary approach. We hope this manuscript has suggested how an harmonious interplay of different research perspectives can be the key to solve such complicated issues and that the future challenges that have emerged from this work will not prove us wrong.

Bibliography

[1] Adams, R.P., MacKay, D.J.: Bayesian online changepoint detection. arXiv preprint arXiv:0710.3742 (2007)

[2] Ambady, N., Rosenthal, R.: Half a minute: Predicting teacher evaluations from thin slices of nonverbal behavior and physical attractiveness. Journal of personality and social psychology 64(3), 431 (1993)

[3] Ambady, N., Bernieri, F.J., Richeson, J.A.: Toward a histology of social behavior: Judgmental accuracy from thin slices of the behavioral stream. In: Advances in experimental social psychology, vol. 32, pp. 201–271. Elsevier (2000)

[4] Ambady, N., Hallahan, M., Conner, B.: Accuracy of judgments of sexual orientation from thin slices of behavior. Journal of personality and social psychology 77(3), 538 (1999)

[5] Ambady, N., Rosenthal, R.: Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. Psychological bulletin 111(2), 256 (1992)

[6] Aminikhanghahi, S., Cook, D.J.: A survey of methods for time series change point detection. Knowledge and information systems 51(2), 339–367 (2017)

[7] Bakeman, R., Gottman, J.M.: Observing interaction: An introduction to sequential analysis. Cambridge university press (1997)

[8] Baldwin, D., Baird, J., Saylor, M., Clark, M.: Infants parse dynamic action. Child development 72(3), 708–717 (2001)

[9] Baldwin, D., Andersson, A., Saffran, J., Meyer, M.: Segmenting dynamic human action via statistical structure. Cognition 106(3), 1382–1407 (2008)

[10] Barbič, J., Safonova, A., Pan, J., Faloutsos, C., Hodgins, J.K., Pollard, N.S.: Segmenting motion capture data into distinct behaviors. In: Proceedings of Graphics Interface 2004. pp. 185–194. Canadian Human-Computer Communications Society (2004)

[11] Borkenau, P., Mauer, N., Riemann, R., Spinath, F.M., Angleitner, A.: Thin slices of behavior as cues of personality and intelligence. Journal of personality and social psychology 86(4), 599 (2004)

[12] Brauner, E., Boos, M., Kolbe, M.E.: The Cambridge Handbook of group interaction analysis. Cambridge University Press. (2018)

[13] Brawley, L., Carron, A., Widmeyer, W.: Assessing the cohesion of teams: Validity of the group environment questionnaire. Journal of Sport Psychology 9(3), 275–294 (1987)

[14] van den Burg, G.J., Williams, C.K.: An evaluation of change point detection algorithms. arXiv preprint arXiv:2003.06222 (2020)

[15] Cabrieto, J., Tuerlinckx, F., Kuppens, P., Grassmann, M., Ceulemans, E.: Detecting correlation changes in multivariate time series: A comparison of four non-parametric change point detection methods. Behavior Research Methods 49(3), 988–1005 (2017)

[16] Camurri, A., Coletta, P., Varni, G., Ghisio, S.: Developing multimodal interactive systems with eyesweb xmi. In: Proceedings of the 7th international conference on New interfaces for musical expression. pp. 305–308 (2007)

[17] Camurri, A., Volpe, G., Piana, S., Mancini, M., Niewiadomski, R., Ferrari, N., Canepa, C.: The dancer in the eye: towards a multi-layered computational framework of qualities in movement. In: Proceedings of the 3rd International Symposium on Movement and Computing. pp. 1–7 (2016)

[18] Carron, A.V., Brawley, L.e.R.: Cohesion: Conceptual and measurement issues. Small group research 31(1), 89–106 (2000)

[19] Ceccaldi, E., Lehmann-Willenbrock, N., Volta, E., Chetouani, M., Volpe, G., Varni, G.: How unitizing affects annotation of cohesion. In: 8th International Conference on Affective Computing Intelligent Interaction. AAAC (2019)

[20] Ceccaldi, E., Volpe, G.: Towards a cognitive-inspired automatic unitizing technique: a feasibility study. In: Proceedings of the International Conference on Advanced Visual Interfaces. pp. 1–5 (2020)

[21] Ceccaldi, E., Volpe, G.: Towards a cognitive-inspired automatic unitizing technique: A feasibility study. In: Proceedings of the International Conference on Advanced Visual Interfaces. AVI '20, Association for Computing Machinery, New York, NY, USA (2020), `https://doi.org/10.1145/3399715.3399825`

[22] Celiktutan, O., Skordos, E., Gunes, H.: Multimodal human-human-robot interactions (mhhri) dataset for studying personality and engagement. IEEE Transactions on Affective Computing 10(4), 484–497 (2017)

[23] Chandola, V., Vatsavai, R.R.: A gaussian process based online change detection algorithm for monitoring periodic time series. In: Proceedings of the 2011 SIAM International Conference on Data Mining. pp. 95–106. SIAM (2011)

[24] Chang, S., Jia, L., Takeuchi, R., Cai, Y.: Do high-commitment work systems affect creativity? A multilevel combinational approach to employee creativity. Journal of Applied Psychology 99(4), 665 (2014)

[25] Chen, J., Gupta, A.K.: Parametric statistical change point analysis: with applications to genetics, medicine, and finance. Springer Science & Business Media (2011)

[26] Chen, L., Harper, M.P.: Multimodal floor control shift detection. In: Proceedings of the 2009 international conference on Multimodal interfaces. pp. 15–22 (2009)

[27] Chin, W., Salisbury, W., Pearson, A.W., Stollak, M.J.: Perceived cohesion in small groups: Adapting and testing the perceived cohesion scale in a small-group setting. Small group research 30(6), 751–766 (1999)

[28] Cruz, A.C., Bhanu, B., Thakoor, N.S.: Vision and attention theory based sampling for continuous facial emotion recognition. IEEE Transactions on Affective Computing 5(4), 418–431 (2014)

[29] Dainton, B.: Temporal consciousness (2010)

[30] Danilo, C., Roberto, B., Lombardo, V., Eleonora, C.: Automatic recognition of narrative drama units: a structured learning approach. In: Text2Story 2019 Second Workshop on Narrative Extraction From Texts. vol. 2342, pp. 81–88. CEUR-WS. org (2019)

[31] Dawes, J.: Do data characteristics change according to the number of scale points used? An experiment using 5-point, 7-point and 10-point scales. International journal of market research 50(1), 61–104 (2008)

[32] Demšar, J.: Statistical comparisons of classifiers over multiple data sets. The Journal of Machine Learning Research 7, 1–30 (2006)

[33] Dennett, D.C., Kinsbourne, M.: Time and the observer: The where and when of consciousness in the brain. Behavioral and Brain sciences 15(2), 183–201 (1992)

[34] Dhall, A., Ramana M., O., Goecke, R., Joshi, J., Gedeon, T.: Video and image based emotion recognition challenges in the wild: Emotiw 2015. In: Proceedings of the 2015 ACM on International Conference on Multimodal Interaction. pp. 423–426. ICMI '15, ACM, New York, NY, USA (2015)

[35] D'Mello, S.K.: On the influence of an iterative affect annotation approach on inter-observer and self-observer reliability. IEEE Transactions on Affective Computing 7(2), 136–149 (2015)

[36] Estabrooks, P.A., Carron, A.V.: The physical activity group environment questionnaire: An instrument for the assessment of cohesion in exercise classes. Group Dynamics: Theory, Research, and Practice 4(3), 230–243 (2000)

[37] Friston, K.: The history of the future of the bayesian brain. NeuroImage 62(2), 1230–1233 (2012)

[38] Gatica-Perez, D., McCowan, L., Zhang, D., Bengio, S.: Detecting group interest-level in meetings. In: Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005. vol. 1, pp. I/489–I/492. IEEE (2005)

[39] Guetzkow, H.: Unitizing and categorizing problems in coding qualitative data. Journal of Clinical Psychology 6(1), 47–58 (1950)

[40] Haynes, K., Fearnhead, P., Eckley, I.A.: A computationally efficient nonparametric approach for changepoint detection. Statistics and Computing 27(5), 1293–1305 (2017)

[41] Hemamou, L., Felhi, G., Martin, J.C., Clavel, C.: Slices of attention in asynchronous video job interviews. In: 2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII). pp. 1–7. IEEE (2019)

[42] Hespos, S.J., Grossman, S.R., Saylor, M.M.: Infants' ability to parse continuous actions: Further evidence. Neural Networks 23(8-9), 1026–1032 (2010)

[43] Hommel, B., Müsseler, J., Aschersleben, G., Prinz, W.: The theory of event coding (tec): A framework for perception and action planning. Behavioral and brain sciences 24(5), 849 (2001)

[44] Hung, H., Gatica-Perez, D.: Estimating cohesion in small groups using audio-visual nonverbal behavior. IEEE Transactions on Multimedia 12(6), 563–575 (2010)

[45] Jakobson, R., Halle, M.: Fundamentals of language, vol. 1. Walter de Gruyter (2010)

[46] James, N.A., Matteson, D.S.: ecp: An r package for nonparametric multiple change point analysis of multivariate data. arXiv preprint arXiv:1309.3295 (2013)

[47] Joo, H., Simon, T., Li, X., Liu, H., Tan, L., Gui, L., Banerjee, S., Godisart, T.S., Nabbe, B., Matthews, I., Kanade, T., Nobuhara, S., Sheikh, Y.: Panoptic studio: A massively multiview system for social interaction capture. IEEE Transactions on Pattern Analysis and Machine Intelligence 41(1), 190–204 (2017)

[48] Kanamori, T., Hido, S., Sugiyama, M.: A least-squares approach to direct importance estimation. The Journal of Machine Learning Research 10, 1391–1445 (2009)

[49] Kantharaju, R.B., Langlet, C., Barange, M., Clavel, C., Pelachaud, C.: Multimodal analysis of cohesion in multi-party interactions. In: LREC (2020)

[50] Kendon, A.: Conducting interaction: Patterns of behavior in focused encounters, vol. 7. CUP Archive (1990)

[51] Keogh, E., Chu, S., Hart, D., Pazzani, M.: Segmenting time series: A survey and novel approach. In: Data mining in time series databases, pp. 1–21. World Scientific (2004)

[52] Keren, S., Gal, A., Karpas, E.: Goal recognition design-survey

[53] Killick, R., Fearnhead, P., Eckley, I.A.: Optimal detection of changepoints with a linear computational cost. Journal of the American Statistical Association 107(500), 1590–1598 (2012)

[54] Klapuri, A.: Sound onset detection by applying psychoacoustic knowledge. In: 1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No. 99CH36258). vol. 6, pp. 3089–3092. IEEE (1999)

[55] Knill, D.C., Pouget, A.: The bayesian brain: the role of uncertainty in neural coding and computation. TRENDS in Neurosciences 27(12), 712–719 (2004)

[56] Koo, T.K., Li, M.Y.: A guideline of selecting and reporting intraclass correlation coefficients for reliability research. Journal of chiropractic medicine 15(2), 155–163 (2016)

[57] Kosie, J., Baldwin, D.: The role of familiarity in segmentation of human action. In: Proceedings of the Annual Meeting of the Cognitive Science Society. vol. 36 (2014)

[58] Kosie, J.E., Baldwin, D.: Attentional profiles linked to event segmentation are robust to missing information. Cognitive research: principles and implications 4(1), 1–18 (2019)

[59] Kurby, C.A., Zacks, J.M.: Segmentation in the perception and memory of events. Trends in cognitive sciences 12(2), 72–79 (2008)

[60] Lehmann-Willenbrock, N., Allen, J.A.: Modeling temporal interaction dynamics in organizational settings. Journal of Business and Psychology 33, 325–344 (2018)

[61] Levine, D., Buchsbaum, D., Hirsh-Pasek, K., Golinkoff, R.: Finding events in a continuous world: A developmental account. Developmental psychobiology 61(3), 376–389 (2018)

[62] Levine, D., Hirsh-Pasek, K., Pace, A., Michnick Golinkoff, R.: A goal bias in action: The boundaries adults perceive in events align with sites of actor intent. Journal of Experimental Psychology: Learning, Memory, and Cognition 43(6), 916 (2017)

[63] Levine, D., Buchsbaum, D., Hirsh-Pasek, K., Golinkoff, R.M.: Finding events in a continuous world: A developmental account. Developmental psychobiology 61(3), 376–389 (2019)

[64] Lin, J.F.S., Karg, M., Kulić, D.: Movement primitive segmentation for human motion modeling: A framework for analysis. IEEE Transactions on Human-Machine Systems 46(3), 325–339 (2016)

[65] Lingenfelser, F., Wagner, J., Deng, J., Brueckner, R., Schuller, B., André, E.: Asynchronous and event-based fusion systems for affect recognition on naturalistic data in comparison to conventional approaches. IEEE Transactions on Affective Computing 9(4), 410–423 (2016)

[66] Liu, S., Yamada, M., Collier, N., Sugiyama, M.: Change-point detection in time-series data by relative density-ratio estimation. Neural Networks 43, 72–83 (2013)

[67] Lombardo, V., Battaglino, C., Pizzo, A., Damiano, R., Lieto, A.: Coupling conceptual modeling and rules for the annotation of dramatic media. Semantic Web 6(5), 503–534 (2015)

[68] Love, J., Selker, R., Marsman, M., Jamil, T., Dropmann, D., Verhagen, J., Ly, A., Gronau, Q.F., Smira, M., Epskamp, S., et al.: Jasp: Graphical statistical software for common statistical designs. Journal of Statistical Software 88(2), 1–17 (2019)

[69] Magee, J.C., Tiedens, L.Z.: Emotional ties that bind: The roles of valence and consistency of group emotion in inferences of cohesiveness and common fate. Personality and Social Psychology Bulletin 32, 1703–1715 (2006)

[70] Maman, L., Ceccaldi, E., Lehmann-Willenbrock, N., Likforman-Sulem, L., Chetouani, M., Volpe, G., Varni, G.: Game-on: A multimodal dataset for cohesion and group analysis. IEEE Access 8, 124185–124203 (2020)

[71] Matteson, D.S.: Nonparametric estimation of change points and stationarity in finance. Change (1/32) (2013)

[72] Matteson, D.S., James, N.A.: A nonparametric approach for multiple change point analysis of multivariate data. Journal of the American Statistical Association 109(505), 334–345 (2014)

[73] Meier, F., Theodorou, E., Stulp, F., Schaal, S.: Movement segmentation using a primitive library. In: 2011 IEEE/RSJ International Conference on Intelligent Robots and Systems. pp. 3407–3412. IEEE (2011)

[74] Meinecke, A.L., Lehmann-Willenbrock, N.: Social Dynamics at Work: Meetings as a gateway, pp. 325—-356. Cambridge Handbooks in Psychology, Cambridge University Press (2015)

[75] Nanninga, M.C., Zhang, Y., Lehmann-Willenbrock, N., Szlávik, Z., Hung, H.: Estimating verbal expressions of task and social cohesion in meetings by quantifying paralinguistic mimicry. In: Proceedings of the 19th ACM International Conference on Multimodal Interaction. pp. 206–215. ICMI '17, ACM, New York, NY, USA (2017)

[76] Newtson, D, a.E.G.: The perceptual organization of ongoing behavior. Journal of Experimental Social Psychology 12(5), 436–450 (1976)

[77] Niewiadomski, R., Hyniewska, S.J., Pelachaud, C.: Constraint-based model for synthesis of multimodal sequential expressions of emotions. IEEE Transactions on Affective Computing 2(3), 134–146 (2011)

[78] Passonneau, R.J., Litman, D.: Discourse segmentation by human and automated means. Computational Linguistics 23(1), 103–139 (1997)

[79] Pearson, R.K., Neuvo, Y., Astola, J., Gabbouj, M.: Generalized hampel filters. EURASIP Journal on Advances in Signal Processing 2016(1), 1–18 (2016)

[80] Pereira, D.G., Afonso, A., Medeiros, F.M.: Overview of friedman's test and post-hoc analysis. Communications in Statistics-Simulation and Computation 44(10), 2636–2653 (2015)

[81] Picard, R.W.: Affective computing mit press. Cambridge, Massachsusetts (1997)

[82] Qu, F., Wang, S.J., Yan, W.J., Li, H., Wu, S., Fu, X.: Cas (me)ˆ2: A database for spontaneous macro-expression and micro-expression spotting and recognition. IEEE Transactions on Affective Computing 9(4), 424–436 (2017)

[83] Reynolds, J.R., Zacks, J.M., Braver, T.S.: A computational model of event segmentation from perceptual prediction. Cognitive science 31(4), 613–643 (2007)

[84] Richeson, J.A., Shelton, J.N.: Brief report: Thin slices of racial bias. Journal of Nonverbal Behavior 29(1), 75–86 (2005)

[85] Rohlfing, K.J., Leonardi, G., Nomikou, I., Raczaszek-Leonardi, J., Hüllermeier, E.: Multimodal turn-taking: motivations, methodological challenges, and novel approaches. IEEE Transactions on Cognitive and Developmental Systems (2019)

[86] Salas, E., Grossman, R., Hughes, A.M., Coultas, C.W.: Measuring team cohesion: Observations from the science. Human factors 57(3), 365–374 (2015)

[87] Samrose, S., Chu, W., He, C., Gao, Y., Shahrin, S.S., Bai, Z., Hoque, M.E.: Visual cues for disrespectful conversation analysis. In: 2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII). pp. 580–586. IEEE (2019)

[88] Schütze, H., Hull, D.A., Pedersen, J.O.: A comparison of classifiers and document representations for the routing problem. In: Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval. pp. 229–237 (1995)

[89] Shaw, R., Cutting, J.: Clues from an ecological theory of event perception. Signed and spoken language: Biological constraints on linguistic form pp. 57–84 (1980)

[90] Shriberg, E., Stolcke, A., Hakkani-Tür, D., Tür, G.: Prosody-based automatic segmentation of speech into sentences and topics. Speech communication 32(1-2), 127–154 (2000)

[91] Skogstad, S., Nymoen, K., Høvin, M., Holm, S., Jensenius, A.: Filtering motion capture data for real-time applications. In: Proceedings of the International Conference on New Interfaces for Musical Expression. pp. 142–147. Graduate School of Culture Technology, KAIST, Daejeon, Republic of Korea (May 2013), http://nime.org/proceedings/2013/nime2013$_2$38*.pdf*

[92] Speer, N.K., Swallow, K.M., Zacks, J.M.: Activation of human motion processing areas during event perception. Cognitive, Affective, & Behavioral Neuroscience 3(4), 335–345 (2003)

[93] Stawarczyk, D., Wahlheim, C.N., Etzel, J.A., Snyder, A.Z., Zacks, J.M.: Aging and the encoding of changes in events: The role of neural activity pattern reinstatement. Proceedings of the National Academy of Sciences 117(47), 29346–29353 (2020)

[94] Treisman, A.: Feature binding, attention and object perception. Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences 353(1373), 1295–1306 (1998)

[95] Van Riemsdijk, M.B., Dastani, M., Winikoff, M.: Goals in agent systems: A unifying framework. In: Proceedings of the 7th international joint conference on Autonomous agents and multiagent systems-Volume 2. pp. 713–720 (2008)

[96] VanRullen, R., Koch, C.: Is perception discrete or continuous? Trends in cognitive sciences 7(5), 207–213 (2003)

[97] Waller, M.J., Kaplan, S.: Systematic behavioral observation for emergent team phenomena: Key considerations for quantitative video-based approaches. Organizational Research Methods 21(2), 500–515 (2018)

[98] Wang, S., Liu, Z., Wang, Z., Wu, G., Shen, P., He, S., Wang, X.: Analyses of a multimodal spontaneous facial expression database. IEEE Transactions on Affective Computing 4(1), 34–46 (2012)

[99] Warkentin, M.E., Sayeed, L., Hightower, R.: Virtual teams versus face-to-face teams: an exploratory study of a web-based conference system. Decision Sciences 28(4), 975–996 (1997)

[100] Weinland, D., Ronfard, R., Boyer, E.: A survey of vision-based methods for action representation, segmentation and recognition. Computer vision and image understanding 115(2), 224–241 (2011)

[101] White, P.A.: Is conscious perception a series of discrete temporal frames? Consciousness and cognition 60, 98–126 (2018)

[102] Whitehill, J., Serpell, Z., Lin, Y.C., Foster, A., Movellan, J.R.: The faces of engagement: Automatic recognition of student engagementfrom facial expressions. IEEE Transactions on Affective Computing 5(1), 86–98 (2014)

[103] Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., Sloetjes, H.: Elan: a professional framework for multimodality research. In: 5th International Conference on Language Resources and Evaluation (LREC 2006). pp. 1556–1559 (2006)

[104] Zacks, J., Speer, N.K., Reynolds, J.R.: Segmentation in reading and film comprehension. Journal of Experimental Psychology: General 138(2), 307–327 (2009)

[105] Zacks, J., Tversky, B.: Event structure in perception and conception. Psychological bulletin 127(1), 3–21 (2001)

[106] Zacks, J.M., Kurby, C.A., Eisenberg, M.L., Haroutunian, N.: Prediction error associated with the perceptual segmentation of naturalistic events. Journal of Cognitive Neuroscience 23(12), 4057–4066 (2011)

[107] Zacks, J.M., Speer, N.K., Swallow, K.M., Braver, T.S., Reynolds, J.R.: Event perception: a mind-brain perspective. Psychological bulletin 133(2), 273 (2007)

[108] Zacks, J.M., Swallow, K.M.: Event segmentation. Current directions in psychological science 16(2), 80–84 (2007)

[109] Zhalehpour, S., Onder, O., Akhtar, Z., Erdem, C.E.: Baum-1: A spontaneous audio-visual face database of affective and mental states. IEEE Transactions on Affective Computing 8(3), 300–313 (2016)

[110] Zheng, Y., Zacks, J.M., Markson, L.: The development of event perception and memory. Cognitive Development 54, 100848 (2020)

VII Appendix

			AUT	RuLIF	CPD	OCP	OCPB	PELT	PELTP
Temporal scale	perscale	Min. value	0.238	0.232	0.259	0.265	0.265	0.258	0.246
		Max. value	0.294	0.303	0.323	0.365	0.360	0.324	0.307
		Mean	0.270	0.258	0.296	0.316	0.308	0.292	0.275
		standard deviation	0.011	0.013	0.013	0.019	0.020	0.015	0.014
	multiscale	Min. value	0.285	0.314	0.326	0.293	0.308	0.318	0.324
		Max. value	0.402	0.416	0.423	0.380	0.414	0.422	0.415
		Mean	0.338	0.365	0.370	0.333	0.361	0.369	0.362
		standard deviation	0.025	0.020	0.020	0.020	0.021	0.021	0.020

Table 34: Metrics for F1-scores for time change category, z-scores

			AUT	RuLIF	CPD	OCP	OCPB	PELT	PELTP
Temporal scale	perscale	Min. value	0.206	0.224	0.212	0.219	0.249	0.260	0.236
		Max. value	0.290	0.317	0.308	0.332	0.332	0.361	0.315
		Mean	0.249	0.278	0.263	0.272	0.288	0.313	0.272
		standard deviation	0.019	0.019	0.020	0.020	0.020	0.020	0.019
	multiscale	Min. value	0.285	0.314	0.326	0.293	0.308	0.318	0.324
		Max. value	0.402	0.416	0.423	0.380	0.414	0.422	0.415
		Mean	0.338	0.365	0.370	0.333	0.361	0.369	0.362
		standard deviation	0.025	0.020	0.020	0.020	0.021	0.021	0.020

Table 35: Metrics for F1-scores for space change category, z-scores

			AUT	RuLIF	CPD	OCP	OCPB	PELT	PELTP
Temporal scale	perscale	Min. value	0.206	0.224	0.212	0.219	0.249	0.260	0.236
		Max. value	0.290	0.317	0.308	0.332	0.332	0.361	0.315
		Mean	0.249	0.278	0.263	0.272	0.288	0.313	0.272
		standard deviation	0.019	0.019	0.020	0.020	0.020	0.020	0.019
	multiscale	Min. value	0.178	0.190	0.209	0.219	0.228	0.249	0.182
		Max. value	0.275	0.283	0.295	0.332	0.319	0.342	0.265
		Mean	0.230	0.243	0.250	0.272	0.272	0.296	0.218
		standard deviation	0.018	0.018	0.019	0.020	0.019	0.020	0.019

Table 36: Metrics for F1-scores for location change category, z-scores

Algorithm	scale	segmentation with conceptual features	Mean	SD	N
AUT	multiscale	no	0.101	0.011	100
		yes	0.290	0.017	100
	perscale	no	0.108	0.012	100
		yes	0.374	0.109	100
CPD	multiscale	no	0.140	0.022	100
		yes	0.675	0.028	100
	perscale	no	0.154	0.019	100
		yes	0.578	0.031	100
OCP	multiscale	no	0.187	0.022	100
		yes	0.698	0.034	100
	perscale	no	0.187	0.023	100
		yes	0.697	0.035	100
OCPB	multiscale	no	0.154	0.022	100
		yes	0.779	0.023	100
	perscale	no	0.180	0.023	100
		yes	0.775	0.020	100
PELT	multiscale	no	0.177	0.025	100
		yes	0.619	0.038	100
	perscale	no	0.148	0.021	100
		yes	0.553	0.055	100
PELTP	multiscale	no	0.158	0.023	100
		yes	0.657	0.029	100
	perscale	no	0.145	0.021	100
		yes	0.529	0.038	100
RuLIF	multiscale	no	0.132	0.018	100
		yes	0.665	0.037	100
	perscale	no	0.147	0.028	100
		yes	0.660	0.026	100

Table 37: Descriptive statistics, validation set

Algorithm	scale	segmentation with conceptual features	Mean	SD	N
AUT	multiscale	no	0.101	0.011	100
		yes	0.290	0.017	100
	perscale	no	0.108	0.012	100
		yes	0.374	0.109	100
CPD	multiscale	no	0.159	0.023	100
		yes	0.680	0.020	100
	perscale	no	0.147	0.018	100
		yes	0.588	0.021	100
OCP	multiscale	no	0.140	0.021	100
		yes	0.686	0.034	100
	perscale	no	0.140	0.022	100
		yes	0.685	0.034	100
OCPB	multiscale	no	0.175	0.024	100
		yes	0.678	0.030	100
	perscale	no	0.177	0.019	100
		yes	0.689	0.033	100
PELT	multiscale	no	0.162	0.022	100
		yes	0.600	0.039	100
	perscale	no	0.152	0.017	100
		yes	0.542	0.036	100
PELTP	multiscale	no	0.144	0.021	100
		yes	0.585	0.038	100
	perscale	no	0.135	0.019	100
		yes	0.520	0.033	100
RuLIF	multiscale	no	0.122	0.017	100
		yes	0.623	0.035	100
	perscale	no	0.129	0.023	100
		yes	0.649	0.027	100

Table 38: Descriptive statistics, validation set, z-scores

Algorithm	scale	segmentation with conceptual features	Mean	SD	N
AUT	multiscale	no	0.116	0.034	95
		yes	0.271	0.070	100
	perscale	no	0.121	0.032	98
		yes	0.387	0.161	100
CPD	multiscale	no	0.135	0.045	95
		yes	0.703	0.058	100
	perscale	no	0.137	0.054	98
		yes	0.580	0.057	100
OCP	multiscale	no	0.188	0.054	95
		yes	0.718	0.054	100
	perscale	no	0.190	0.052	98
		yes	0.716	0.053	100
OCPB	multiscale	no	0.130	0.044	95
		yes	0.789	0.037	100
	perscale	no	0.170	0.066	98
		yes	0.801	0.042	100
PELT	multiscale	no	0.150	0.054	95
		yes	0.640	0.057	100
	perscale	no	0.135	0.052	98
		yes	0.584	0.067	100
PELTP	multiscale	no	0.130	0.058	95
		yes	0.677	0.061	100
	perscale	no	0.134	0.056	98
		yes	0.552	0.080	100
RuLIF	multiscale	no	0.126	0.044	95
		yes	0.685	0.068	100
	perscale	no	0.115	0.048	98
		yes	0.677	0.061	100

Table 39: Descriptive statistics, test set

Algorithm	**scale**	**segmentation with conceptual features**	**Mean**	**SD**	**N**
AUT	no	multiscale	0.118	0.035	94
		perscale	0.121	0.032	99
	yes	multiscale	0.271	0.070	100
		perscale	0.387	0.161	100
CPD	no	multiscale	0.159	0.063	94
		perscale	0.136	0.046	99
	yes	multiscale	0.699	0.052	100
		perscale	0.579	0.058	100
OCP	no	multiscale	0.118	0.041	94
		perscale	0.119	0.039	99
	yes	multiscale	0.707	0.056	100
		perscale	0.707	0.056	100
OCPB	no	multiscale	0.138	0.054	94
		perscale	0.135	0.049	99
	yes	multiscale	0.620	0.060	100
		perscale	0.569	0.058	100
PELT	no	multiscale	0.133	0.054	94
		perscale	0.118	0.053	99
	yes	multiscale	0.619	0.074	100
		perscale	0.546	0.078	100
PELTP	no	multiscale	0.131	0.059	94
		perscale	0.134	0.056	99
	yes	multiscale	0.677	0.061	100
		perscale	0.552	0.080	100
RuLIF	no	multiscale	0.104	0.040	94
		perscale	0.103	0.045	99
	yes	multiscale	0.632	0.044	100
		perscale	0.668	0.070	100

Table 40: Descriptive statistics, test set, z-scores

		T-Stat	df	W_i	W_j	p	p_{bonf}	p_{holm}
AUT	RuLIF	5.738	594	103.000	278.500	$< .001$	$< .001$	$< .001$
	CPD	7.014	594	103.000	317.500	$< .001$	$< .001$	$< .001$
	OCP	17.477	594	103.000	637.500	$< .001$	$< .001$	$< .001$
	OCPB	10.038	594	103.000	410.000	$< .001$	$< .001$	$< .001$
	PELT	15.973	594	103.000	591.500	$< .001$	$< .001$	$< .001$
	PELTP	11.738	594	103.000	462.000	$< .001$	$< .001$	$< .001$
RuLIF	CPD	1.275	594	278.500	317.500	0.203	1.000	0.269
	OCP	11.738	594	278.500	637.500	$< .001$	$< .001$	$< .001$
	OCPB	4.300	594	278.500	410.000	$< .001$	$< .001$	$< .001$
	PELT	10.234	594	278.500	591.500	$< .001$	$< .001$	$< .001$
	PELTP	6.000	594	278.500	462.000	$< .001$	$< .001$	$< .001$
CPD	OCP	10.463	594	317.500	637.500	$< .001$	$< .001$	$< .001$
	OCPB	3.025	594	317.500	410.000	0.003	0.055	0.010
	PELT	8.959	594	317.500	591.500	$< .001$	$< .001$	$< .001$
	PELTP	4.725	594	317.500	462.000	$< .001$	$< .001$	$< .001$
OCP	OCPB	7.439	594	637.500	410.000	$< .001$	$< .001$	$< .001$
	PELT	1.504	594	637.500	591.500	0.133	1.000	0.269
	PELTP	5.738	594	637.500	462.000	$< .001$	$< .001$	$< .001$
OCPB	PELT	5.935	594	410.000	591.500	$< .001$	$< .001$	$< .001$
	PELTP	1.700	594	410.000	462.000	0.090	1.000	0.269
PELT	PELTP	4.234	594	591.500	462.000	$< .001$	$< .001$	$< .001$

Table 41: Conover's Post Hoc Comparisons for multiscale segmentations on validation sets without conceptual features, scores

		T-Stat	df	W_i	W_j	p	p_{bonf}	p_{holm}
AUT	RuLIF	4.251	594	114.000	244.000	$< .001$	$< .001$	$< .001$
	CPD	13.245	594	114.000	519.000	$< .001$	$< .001$	$< .001$
	OCP	8.437	594	114.000	372.000	$< .001$	$< .001$	$< .001$
	OCPB	17.055	594	114.000	635.500	$< .001$	$< .001$	$< .001$
	PELT	9.173	594	114.000	394.500	$< .001$	$< .001$	$< .001$
	PELTP	13.310	594	114.000	521.000	$< .001$	$< .001$	$< .001$
RuLIF	CPD	8.993	594	244.000	519.000	$< .001$	$< .001$	$< .001$
	OCP	4.186	594	244.000	372.000	$< .001$	$< .001$	$< .001$
	OCPB	12.803	594	244.000	635.500	$< .001$	$< .001$	$< .001$
	PELT	4.922	594	244.000	394.500	$< .001$	$< .001$	$< .001$
	PELTP	9.059	594	244.000	521.000	$< .001$	$< .001$	$< .001$
CPD	OCP	4.807	594	519.000	372.000	$< .001$	$< .001$	$< .001$
	OCPB	3.810	594	519.000	635.500	$< .001$	0.003	$< .001$
	PELT	4.072	594	519.000	394.500	$< .001$	0.001	$< .001$
	PELTP	0.065	594	519.000	521.000	0.948	1.000	0.948
OCP	OCPB	8.617	594	372.000	635.500	$< .001$	$< .001$	$< .001$
	PELT	0.736	594	372.000	394.500	0.462	1.000	0.924
	PELTP	4.873	594	372.000	521.000	$< .001$	$< .001$	$< .001$
OCPB	PELT	7.881	594	635.500	394.500	$< .001$	$< .001$	$< .001$
	PELTP	3.745	594	635.500	521.000	$< .001$	0.004	$< .001$
PELT	PELTP	4.137	594	394.500	521.000	$< .001$	$< .001$	$< .001$

Table 42: Conover's Post Hoc Comparisons for multiscale segmentations on validation sets without conceptual features, z-scores

		T-Stat	df	W_i	W_j	p	p_{bonf}	p_{holm}
AUT	RuLIF	8.203	594	103.500	354.500	$< .001$	$< .001$	$< .001$
	CPD	10.245	594	103.500	417.000	$< .001$	$< .001$	$< .001$
	OCP	17.484	594	103.500	638.500	$< .001$	$< .001$	$< .001$
	OCPB	15.572	594	103.500	580.000	$< .001$	$< .001$	$< .001$
	PELT	8.660	594	103.500	368.500	$< .001$	$< .001$	$< .001$
	PELTP	7.663	594	103.500	338.000	$< .001$	$< .001$	$< .001$
RuLIF	CPD	2.042	594	354.500	417.000	0.042	0.872	0.249
	OCP	9.281	594	354.500	638.500	$< .001$	$< .001$	$< .001$
	OCPB	7.369	594	354.500	580.000	$< .001$	$< .001$	$< .001$
	PELT	0.458	594	354.500	368.500	0.647	1.000	1.000
	PELTP	0.539	594	354.500	338.000	0.590	1.000	1.000
CPD	OCP	7.239	594	417.000	638.500	$< .001$	$< .001$	$< .001$
	OCPB	5.327	594	417.000	580.000	$< .001$	$< .001$	$< .001$
	PELT	1.585	594	417.000	368.500	0.114	1.000	0.454
	PELTP	2.582	594	417.000	338.000	0.010	0.211	0.070
OCP	OCPB	1.912	594	638.500	580.000	0.056	1.000	0.282
	PELT	8.824	594	638.500	368.500	$< .001$	$< .001$	$< .001$
	PELTP	9.820	594	638.500	338.000	$< .001$	$< .001$	$< .001$
OCPB	PELT	6.912	594	580.000	368.500	$< .001$	$< .001$	$< .001$
	PELTP	7.908	594	580.000	338.000	$< .001$	$< .001$	$< .001$
PELT	PELTP	0.997	594	368.500	338.000	0.319	1.000	0.958

Table 43: Conover's Post Hoc Comparisons for perscale segmentations on validation sets without conceptual features, scores

		T-Stat	df	W_i	W_j	p	p_{bonf}	p_{holm}
AUT	RuLIF	5.546	594	117.500	287.000	$< .001$	$< .001$	$< .001$
	CPD	11.747	594	117.500	476.500	$< .001$	$< .001$	$< .001$
	OCP	8.966	594	117.500	391.500	$< .001$	$< .001$	$< .001$
	OCPB	17.915	594	117.500	665.000	$< .001$	$< .001$	$< .001$
	PELT	13.056	594	117.500	516.500	$< .001$	$< .001$	$< .001$
	PELTP	7.477	594	117.500	346.000	$< .001$	$< .001$	$< .001$
RuLIF	CPD	6.201	594	287.000	476.500	$< .001$	$< .001$	$< .001$
	OCP	3.419	594	287.000	391.500	$< .001$	0.014	0.003
	OCPB	12.368	594	287.000	665.000	$< .001$	$< .001$	$< .001$
	PELT	7.509	594	287.000	516.500	$< .001$	$< .001$	$< .001$
	PELTP	1.931	594	287.000	346.000	0.054	1.000	0.162
CPD	OCP	2.781	594	476.500	391.500	0.006	0.117	0.022
	OCPB	6.168	594	476.500	665.000	$< .001$	$< .001$	$< .001$
	PELT	1.309	594	476.500	516.500	0.191	1.000	0.274
	PELTP	4.270	594	476.500	346.000	$< .001$	$< .001$	$< .001$
OCP	OCPB	8.949	594	391.500	665.000	$< .001$	$< .001$	$< .001$
	PELT	4.090	594	391.500	516.500	$< .001$	0.001	$< .001$
	PELTP	1.489	594	391.500	346.000	0.137	1.000	0.274
OCPB	PELT	4.859	594	665.000	516.500	$< .001$	$< .001$	$< .001$
	PELTP	10.438	594	665.000	346.000	$< .001$	$< .001$	$< .001$
PELT	PELTP	5.579	594	516.500	346.000	$< .001$	$< .001$	$< .001$

Table 44: Conover's Post Hoc Comparisons for perscale segmentations on validation sets without conceptual features, z-scores

		T-Stat	df	W_i	W_j	p	p_{bonf}	p_{holm}
AUT	RuLIF	12.427	594	113.500	494.000	$< .001$	$< .001$	$< .001$
	CPD	8.018	594	113.500	359.000	$< .001$	$< .001$	$< .001$
	OCP	15.611	594	113.500	591.500	$< .001$	$< .001$	$< .001$
	OCPB	19.089	594	113.500	698.000	$< .001$	$< .001$	$< .001$
	PELT	6.042	594	113.500	298.500	$< .001$	$< .001$	$< .001$
	PELTP	4.311	594	113.500	245.500	$< .001$	$< .001$	$< .001$
RuLIF	CPD	4.409	594	494.000	359.000	$< .001$	$< .001$	$< .001$
	OCP	3.184	594	494.000	591.500	0.002	0.032	0.005
	OCPB	6.662	594	494.000	698.000	$< .001$	$< .001$	$< .001$
	PELT	6.385	594	494.000	298.500	$< .001$	$< .001$	$< .001$
	PELTP	8.116	594	494.000	245.500	$< .001$	$< .001$	$< .001$
CPD	OCP	7.593	594	359.000	591.500	$< .001$	$< .001$	$< .001$
	OCPB	11.071	594	359.000	698.000	$< .001$	$< .001$	$< .001$
	PELT	1.976	594	359.000	298.500	0.049	1.000	0.097
	PELTP	3.707	594	359.000	245.500	$< .001$	0.005	0.001
OCP	OCPB	3.478	594	591.500	698.000	$< .001$	0.011	0.002
	PELT	9.569	594	591.500	298.500	$< .001$	$< .001$	$< .001$
	PELTP	11.300	594	591.500	245.500	$< .001$	$< .001$	$< .001$
OCPB	PELT	13.047	594	698.000	298.500	$< .001$	$< .001$	$< .001$
	PELTP	14.778	594	698.000	245.500	$< .001$	$< .001$	$< .001$
PELT	PELTP	1.731	594	298.500	245.500	0.084	1.000	0.097

Table 45: Conover's Post Hoc Comparisons for perscale segmentations on validation sets with conceptual features, scores

		T-Stat	df	W_i	W_j	p	p_{bonf}	p_{holm}
AUT	RuLIF	13.453	594	116.000	528.000	$< .001$	$< .001$	$< .001$
	CPD	8.963	594	116.000	390.500	$< .001$	$< .001$	$< .001$
	OCP	16.457	594	116.000	620.000	$< .001$	$< .001$	$< .001$
	OCPB	17.241	594	116.000	644.000	$< .001$	$< .001$	$< .001$
	PELT	5.159	594	116.000	274.000	$< .001$	$< .001$	$< .001$
	PELTP	3.641	594	116.000	227.500	$< .001$	0.006	0.001
RuLIF	CPD	4.490	594	528.000	390.500	$< .001$	$< .001$	$< .001$
	OCP	3.004	594	528.000	620.000	0.003	0.058	0.008
	OCPB	3.788	594	528.000	644.000	$< .001$	0.004	$< .001$
	PELT	8.294	594	528.000	274.000	$< .001$	$< .001$	$< .001$
	PELTP	9.812	594	528.000	227.500	$< .001$	$< .001$	$< .001$
CPD	OCP	7.494	594	390.500	620.000	$< .001$	$< .001$	$< .001$
	OCPB	8.278	594	390.500	644.000	$< .001$	$< .001$	$< .001$
	PELT	3.804	594	390.500	274.000	$< .001$	0.003	$< .001$
	PELTP	5.322	594	390.500	227.500	$< .001$	$< .001$	$< .001$
OCP	OCPB	0.784	594	620.000	644.000	0.434	1.000	0.434
	PELT	11.298	594	620.000	274.000	$< .001$	$< .001$	$< .001$
	PELTP	12.816	594	620.000	227.500	$< .001$	$< .001$	$< .001$
OCPB	PELT	12.082	594	644.000	274.000	$< .001$	$< .001$	$< .001$
	PELTP	13.600	594	644.000	227.500	$< .001$	$< .001$	$< .001$
PELT	PELTP	1.518	594	274.000	227.500	0.129	1.000	0.259

Table 46: Conover's Post Hoc Comparisons for perscale segmentations on validation sets with conceptual features, z-scores

		T-Stat	**df**	**W**$_i$	**W**$_j$	**p**	**p**$_{bonf}$	**p**$_{holm}$
AUT	RuLIF	11.052	594	100.000	438.500	$< .001$	$< .001$	$< .001$
	CPD	12.407	594	100.000	480.000	$< .001$	$< .001$	$< .001$
	OCP	14.562	594	100.000	546.000	$< .001$	$< .001$	$< .001$
	OCPB	19.558	594	100.000	699.000	$< .001$	$< .001$	$< .001$
	PELT	5.942	594	100.000	282.000	$< .001$	$< .001$	$< .001$
	PELTP	5.044	594	100.000	254.500	$< .001$	$< .001$	$< .001$
RuLIF	CPD	1.355	594	438.500	480.000	0.176	1.000	0.352
	OCP	3.510	594	438.500	546.000	$< .001$	0.010	0.002
	OCPB	8.505	594	438.500	699.000	$< .001$	$< .001$	$< .001$
	PELT	5.110	594	438.500	282.000	$< .001$	$< .001$	$< .001$
	PELTP	6.008	594	438.500	254.500	$< .001$	$< .001$	$< .001$
CPD	OCP	2.155	594	480.000	546.000	0.032	0.663	0.095
	OCPB	7.150	594	480.000	699.000	$< .001$	$< .001$	$< .001$
	PELT	6.465	594	480.000	282.000	$< .001$	$< .001$	$< .001$
	PELTP	7.363	594	480.000	254.500	$< .001$	$< .001$	$< .001$
OCP	OCPB	4.996	594	546.000	699.000	$< .001$	$< .001$	$< .001$
	PELT	8.620	594	546.000	282.000	$< .001$	$< .001$	$< .001$
	PELTP	9.518	594	546.000	254.500	$< .001$	$< .001$	$< .001$
OCPB	PELT	13.615	594	699.000	282.000	$< .001$	$< .001$	$< .001$
	PELTP	14.513	594	699.000	254.500	$< .001$	$< .001$	$< .001$
PELT	PELTP	0.898	594	282.000	254.500	0.370	1.000	0.370

Table 47: Conover's Post Hoc Comparisons for multiscale segmentation on validation sets with conceptual features, scores

		T-Stat	df	W_i	W_j	p	p_{bonf}	p_{holm}
AUT	RuLIF	6.501	594	100.000	299.000	$< .001$	$< .001$	$< .001$
	CPD	15.125	594	100.000	563.000	$< .001$	$< .001$	$< .001$
	OCP	15.811	594	100.000	584.000	$< .001$	$< .001$	$< .001$
	OCPB	14.096	594	100.000	531.500	$< .001$	$< .001$	$< .001$
	PELT	6.289	594	100.000	292.500	$< .001$	$< .001$	$< .001$
	PELTP	10.780	594	100.000	430.000	$< .001$	$< .001$	$< .001$
RuLIF	CPD	8.624	594	299.000	563.000	$< .001$	$< .001$	$< .001$
	OCP	9.310	594	299.000	584.000	$< .001$	$< .001$	$< .001$
	OCPB	7.595	594	299.000	531.500	$< .001$	$< .001$	$< .001$
	PELT	0.212	594	299.000	292.500	0.832	1.000	0.986
	PELTP	4.279	594	299.000	430.000	$< .001$	$< .001$	$< .001$
CPD	OCP	0.686	594	563.000	584.000	0.493	1.000	0.986
	OCPB	1.029	594	563.000	531.500	0.304	1.000	0.912
	PELT	8.837	594	563.000	292.500	$< .001$	$< .001$	$< .001$
	PELTP	4.345	594	563.000	430.000	$< .001$	$< .001$	$< .001$
OCP	OCPB	1.715	594	584.000	531.500	0.087	1.000	0.347
	PELT	9.523	594	584.000	292.500	$< .001$	$< .001$	$< .001$
	PELTP	5.031	594	584.000	430.000	$< .001$	$< .001$	$< .001$
OCPB	PELT	7.808	594	531.500	292.500	$< .001$	$< .001$	$< .001$
	PELTP	3.316	594	531.500	430.000	$< .001$	0.020	0.005
PELT	PELTP	4.492	594	292.500	430.000	$< .001$	$< .001$	$< .001$

Table 48: Conover's Post Hoc Comparisons for multiscale segmentation on validation sets with conceptual features, z-scores

		T-Stat	**df**	**W**$_i$	**W**$_j$	**p**	**p**$_{bonf}$	**p**$_{holm}$
AUT	RuLIF	0.198	582	286.000	292.000	0.843	1.000	1.000
	CPD	3.150	582	286.000	381.500	0.002	0.036	0.019
	OCP	9.666	582	286.000	579.000	$< .001$	$< .001$	$< .001$
	OCPB	6.218	582	286.000	474.500	$< .001$	$< .001$	$< .001$
	PELT	3.249	582	286.000	384.500	0.001	0.026	0.015
	PELTP	1.996	582	286.000	346.500	0.046	0.975	0.278
RuLIF	CPD	2.953	582	292.000	381.500	0.003	0.069	0.025
	OCP	9.468	582	292.000	579.000	$< .001$	$< .001$	$< .001$
	OCPB	6.021	582	292.000	474.500	$< .001$	$< .001$	$< .001$
	PELT	3.051	582	292.000	384.500	0.002	0.050	0.023
	PELTP	1.798	582	292.000	346.500	0.073	1.000	0.364
CPD	OCP	6.515	582	381.500	579.000	$< .001$	$< .001$	$< .001$
	OCPB	3.068	582	381.500	474.500	0.002	0.047	0.023
	PELT	0.099	582	381.500	384.500	0.921	1.000	1.000
	PELTP	1.155	582	381.500	346.500	0.249	1.000	0.842
OCP	OCPB	3.447	582	579.000	474.500	$< .001$	0.013	0.008
	PELT	6.416	582	579.000	384.500	$< .001$	$< .001$	$< .001$
	PELTP	7.670	582	579.000	346.500	$< .001$	$< .001$	$< .001$
OCPB	PELT	2.969	582	474.500	384.500	0.003	0.065	0.025
	PELTP	4.223	582	474.500	346.500	$< .001$	$< .001$	$< .001$
PELT	PELTP	1.254	582	384.500	346.500	0.210	1.000	0.842

Table 49: Conover's Post Hoc Comparisons for perscale segmentation on test sets without conceptual features, scores

		T-Stat	df	W_i	W_j	p	p_{bonf}	p_{holm}
AUT	RuLIF	3.238	588	350.000	251.500	0.001	0.027	0.010
	CPD	4.110	588	350.000	475.000	$< .001$	$< .001$	$< .001$
	OCP	0.164	588	350.000	355.000	0.869	1.000	1.000
	OCPB	6.707	588	350.000	554.000	$< .001$	$< .001$	$< .001$
	PELT	3.271	588	350.000	449.500	0.001	0.024	0.010
	PELTP	0.427	588	350.000	337.000	0.669	1.000	1.000
RuLIF	CPD	7.348	588	251.500	475.000	$< .001$	$< .001$	$< .001$
	OCP	3.403	588	251.500	355.000	$< .001$	0.015	0.007
	OCPB	9.945	588	251.500	554.000	$< .001$	$< .001$	$< .001$
	PELT	6.510	588	251.500	449.500	$< .001$	$< .001$	$< .001$
	PELTP	2.811	588	251.500	337.000	0.005	0.107	0.031
CPD	OCP	3.945	588	475.000	355.000	$< .001$	0.002	0.001
	OCPB	2.597	588	475.000	554.000	0.010	0.202	0.048
	PELT	0.838	588	475.000	449.500	0.402	1.000	1.000
	PELTP	4.537	588	475.000	337.000	$< .001$	$< .001$	$< .001$
OCP	OCPB	6.542	588	355.000	554.000	$< .001$	$< .001$	$< .001$
	PELT	3.107	588	355.000	449.500	0.002	0.042	0.014
	PELTP	0.592	588	355.000	337.000	0.554	1.000	1.000
OCPB	PELT	3.436	588	554.000	449.500	$< .001$	0.013	0.007
	PELTP	7.134	588	554.000	337.000	$< .001$	$< .001$	$< .001$
PELT	PELTP	3.699	588	449.500	337.000	$< .001$	0.005	0.003

Table 50: Conover's Post Hoc Comparisons for perscale segmentation on test sets without conceptual features, scores

		T-Stat	df	W_i	W_j	p	p_{bonf}	p_{holm}
AUT	RuLIF	14.502	1194	272.000	899.000	$< .001$	$< .001$	$< .001$
	CPD	11.749	1194	272.000	780.000	$< .001$	$< .001$	$< .001$
	OCP	18.341	1194	272.000	1065.000	$< .001$	$< .001$	$< .001$
	OCPB	25.280	1194	272.000	1365.000	$< .001$	$< .001$	$< .001$
	PELT	7.135	1194	272.000	580.500	$< .001$	$< .001$	$< .001$
	PELTP	8.477	1194	272.000	638.500	$< .001$	$< .001$	$< .001$
RuLIF	CPD	2.752	1194	899.000	780.000	0.006	0.126	0.012
	OCP	3.839	1194	899.000	1065.000	$< .001$	0.003	$< .001$
	OCPB	10.778	1194	899.000	1365.000	$< .001$	$< .001$	$< .001$
	PELT	7.367	1194	899.000	580.500	$< .001$	$< .001$	$< .001$
	PELTP	6.025	1194	899.000	638.500	$< .001$	$< .001$	$< .001$
CPD	OCP	6.592	1194	780.000	1065.000	$< .001$	$< .001$	$< .001$
	OCPB	13.530	1194	780.000	1365.000	$< .001$	$< .001$	$< .001$
	PELT	4.614	1194	780.000	580.500	$< .001$	$< .001$	$< .001$
	PELTP	3.273	1194	780.000	638.500	0.001	0.023	0.003
OCP	OCPB	6.939	1194	1065.000	1365.000	$< .001$	$< .001$	$< .001$
	PELT	11.206	1194	1065.000	580.500	$< .001$	$< .001$	$< .001$
	PELTP	9.864	1194	1065.000	638.500	$< .001$	$< .001$	$< .001$
OCPB	PELT	18.145	1194	1365.000	580.500	$< .001$	$< .001$	$< .001$
	PELTP	16.803	1194	1365.000	638.500	$< .001$	$< .001$	$< .001$
PELT	PELTP	1.341	1194	580.500	638.500	0.180	1.000	0.180

Table 51: Conover's Post Hoc Comparisons for perscale segmentation on test sets with conceptual features, scores

		T-Stat	df	W_i	W_j	p	p_{bonf}	p_{holm}
AUT	RuLIF	14.113	1194	282.000	892.000	$< .001$	$< .001$	$< .001$
	CPD	14.333	1194	282.000	901.500	$< .001$	$< .001$	$< .001$
	OCP	20.915	1194	282.000	1186.000	$< .001$	$< .001$	$< .001$
	OCPB	20.036	1194	282.000	1148.000	$< .001$	$< .001$	$< .001$
	PELT	7.704	1194	282.000	615.000	$< .001$	$< .001$	$< .001$
	PELTP	6.790	1194	282.000	575.500	$< .001$	$< .001$	$< .001$
RuLIF	CPD	0.220	1194	892.000	901.500	0.826	1.000	1.000
	OCP	6.802	1194	892.000	1186.000	$< .001$	$< .001$	$< .001$
	OCPB	5.923	1194	892.000	1148.000	$< .001$	$< .001$	$< .001$
	PELT	6.409	1194	892.000	615.000	$< .001$	$< .001$	$< .001$
	PELTP	7.323	1194	892.000	575.500	$< .001$	$< .001$	$< .001$
CPD	OCP	6.582	1194	901.500	1186.000	$< .001$	$< .001$	$< .001$
	OCPB	5.703	1194	901.500	1148.000	$< .001$	$< .001$	$< .001$
	PELT	6.629	1194	901.500	615.000	$< .001$	$< .001$	$< .001$
	PELTP	7.542	1194	901.500	575.500	$< .001$	$< .001$	$< .001$
OCP	OCPB	0.879	1194	1186.000	1148.000	0.379	1.000	1.000
	PELT	13.211	1194	1186.000	615.000	$< .001$	$< .001$	$< .001$
	PELTP	14.125	1194	1186.000	575.500	$< .001$	$< .001$	$< .001$
OCPB	PELT	12.332	1194	1148.000	615.000	$< .001$	$< .001$	$< .001$
	PELTP	13.245	1194	1148.000	575.500	$< .001$	$< .001$	$< .001$
PELT	PELTP	0.914	1194	615.000	575.500	0.361	1.000	1.000

Table 52: Conover's Post Hoc Comparisons for perscale segmentation on test sets with conceptual features, z-scores

		T-Stat	df	W_i	W_j	p	p_{bonf}	p_{holm}
AUT	RuLIF	2.631	564	252.000	330.500	0.009	0.184	0.079
	CPD	3.988	564	252.000	371.000	$< .001$	0.002	0.001
	OCP	10.405	564	252.000	562.500	$< .001$	$< .001$	$< .001$
	OCPB	3.753	564	252.000	364.000	$< .001$	0.004	0.003
	PELT	6.199	564	252.000	437.000	$< .001$	$< .001$	$< .001$
	PELTP	3.049	564	252.000	343.000	0.002	0.050	0.024
RuLIF	CPD	1.357	564	330.500	371.000	0.175	1.000	1.000
	OCP	7.774	564	330.500	562.500	$< .001$	$< .001$	$< .001$
	OCPB	1.123	564	330.500	364.000	0.262	1.000	1.000
	PELT	3.569	564	330.500	437.000	$< .001$	0.008	0.005
	PELTP	0.419	564	330.500	343.000	0.675	1.000	1.000
CPD	OCP	6.417	564	371.000	562.500	$< .001$	$< .001$	$< .001$
	OCPB	0.235	564	371.000	364.000	0.815	1.000	1.000
	PELT	2.212	564	371.000	437.000	0.027	0.575	0.192
	PELTP	0.938	564	371.000	343.000	0.349	1.000	1.000
OCP	OCPB	6.652	564	562.500	364.000	$< .001$	$< .001$	$< .001$
	PELT	4.205	564	562.500	437.000	$< .001$	$< .001$	$< .001$
	PELTP	7.355	564	562.500	343.000	$< .001$	$< .001$	$< .001$
OCPB	PELT	2.446	564	364.000	437.000	0.015	0.310	0.118
	PELTP	0.704	564	364.000	343.000	0.482	1.000	1.000
PELT	PELTP	3.150	564	437.000	343.000	0.002	0.036	0.019

Table 53: Conover's Post Hoc Comparisons for multiscale segmentation on test sets without conceptual features, scores

		T-Stat	**df**	$\mathbf{W}_i$	$\mathbf{W}_j$	**p**	$\mathbf{p}_{bonf}$	$\mathbf{p}_{holm}$
AUT	RuLIF	1.046	558	291.000	260.000	0.296	1.000	0.953
	CPD	7.051	558	291.000	500.000	$< .001$	$< .001$	$< .001$
	OCP	0.978	558	291.000	320.000	0.328	1.000	0.953
	OCPB	5.870	558	291.000	465.000	$< .001$	$< .001$	$< .001$
	PELT	3.711	558	291.000	401.000	$< .001$	0.005	0.003
	PELTP	3.508	558	291.000	395.000	$< .001$	0.010	0.005
RuLIF	CPD	8.096	558	260.000	500.000	$< .001$	$< .001$	$< .001$
	OCP	2.024	558	260.000	320.000	0.043	0.912	0.217
	OCPB	6.916	558	260.000	465.000	$< .001$	$< .001$	$< .001$
	PELT	4.757	558	260.000	401.000	$< .001$	$< .001$	$< .001$
	PELTP	4.554	558	260.000	395.000	$< .001$	$< .001$	$< .001$
CPD	OCP	6.072	558	500.000	320.000	$< .001$	$< .001$	$< .001$
	OCPB	1.181	558	500.000	465.000	0.238	1.000	0.953
	PELT	3.340	558	500.000	401.000	$< .001$	0.019	0.009
	PELTP	3.542	558	500.000	395.000	$< .001$	0.009	0.005
OCP	OCPB	4.892	558	320.000	465.000	$< .001$	$< .001$	$< .001$
	PELT	2.733	558	320.000	401.000	0.006	0.136	0.058
	PELTP	2.530	558	320.000	395.000	0.012	0.245	0.093
OCPB	PELT	2.159	558	465.000	401.000	0.031	0.657	0.188
	PELTP	2.361	558	465.000	395.000	0.019	0.389	0.130
PELT	PELTP	0.202	558	401.000	395.000	0.840	1.000	0.953

Table 54: Conover's Post Hoc Comparisons for multiscale segmentation on test sets without conceptual features, z-scores

		T-Stat	df	W_i	W_j	p	p_{bonf}	p_{holm}
AUT	RuLIF	9.587	594	100.000	393.500	$< .001$	$< .001$	$< .001$
	CPD	12.331	594	100.000	477.500	$< .001$	$< .001$	$< .001$
	OCP	12.772	594	100.000	491.000	$< .001$	$< .001$	$< .001$
	OCPB	18.782	594	100.000	675.000	$< .001$	$< .001$	$< .001$
	PELT	5.618	594	100.000	272.000	$< .001$	$< .001$	$< .001$
	PELTP	9.506	594	100.000	391.000	$< .001$	$< .001$	$< .001$
RuLIF	CPD	2.744	594	393.500	477.500	0.006	0.131	0.020
	OCP	3.185	594	393.500	491.000	0.002	0.032	0.008
	OCPB	9.195	594	393.500	675.000	$< .001$	$< .001$	$< .001$
	PELT	3.969	594	393.500	272.000	$< .001$	0.002	$< .001$
	PELTP	0.082	594	393.500	391.000	0.935	1.000	1.000
CPD	OCP	0.441	594	477.500	491.000	0.659	1.000	1.000
	OCPB	6.451	594	477.500	675.000	$< .001$	$< .001$	$< .001$
	PELT	6.713	594	477.500	272.000	$< .001$	$< .001$	$< .001$
	PELTP	2.826	594	477.500	391.000	0.005	0.102	0.020
OCP	OCPB	6.010	594	491.000	675.000	$< .001$	$< .001$	$< .001$
	PELT	7.154	594	491.000	272.000	$< .001$	$< .001$	$< .001$
	PELTP	3.266	594	491.000	391.000	0.001	0.024	0.007
OCPB	PELT	13.164	594	675.000	272.000	$< .001$	$< .001$	$< .001$
	PELTP	9.277	594	675.000	391.000	$< .001$	$< .001$	$< .001$
PELT	PELTP	3.887	594	272.000	391.000	$< .001$	0.002	$< .001$

Table 55: Conover's Post Hoc Comparisons for multiscale segmentation on test sets with conceptual features, scores

		T-Stat	df	W_i	W_j	p	p_{bonf}	p_{holm}
AUT	RuLIF	8.497	594	100.000	360.000	$< .001$	$< .001$	$< .001$
	CPD	15.556	594	100.000	576.000	$< .001$	$< .001$	$< .001$
	OCP	15.490	594	100.000	574.000	$< .001$	$< .001$	$< .001$
	OCPB	14.526	594	100.000	544.500	$< .001$	$< .001$	$< .001$
	PELT	7.091	594	100.000	317.000	$< .001$	$< .001$	$< .001$
	PELTP	7.467	594	100.000	328.500	$< .001$	$< .001$	$< .001$
RuLIF	CPD	7.059	594	360.000	576.000	$< .001$	$< .001$	$< .001$
	OCP	6.993	594	360.000	574.000	$< .001$	$< .001$	$< .001$
	OCPB	6.029	594	360.000	544.500	$< .001$	$< .001$	$< .001$
	PELT	1.405	594	360.000	317.000	0.160	1.000	0.963
	PELTP	1.029	594	360.000	328.500	0.304	1.000	1.000
CPD	OCP	0.065	594	576.000	574.000	0.948	1.000	1.000
	OCPB	1.029	594	576.000	544.500	0.304	1.000	1.000
	PELT	8.464	594	576.000	317.000	$< .001$	$< .001$	$< .001$
	PELTP	8.088	594	576.000	328.500	$< .001$	$< .001$	$< .001$
OCP	OCPB	0.964	594	574.000	544.500	0.335	1.000	1.000
	PELT	8.399	594	574.000	317.000	$< .001$	$< .001$	$< .001$
	PELTP	8.023	594	574.000	328.500	$< .001$	$< .001$	$< .001$
OCPB	PELT	7.435	594	544.500	317.000	$< .001$	$< .001$	$< .001$
	PELTP	7.059	594	544.500	328.500	$< .001$	$< .001$	$< .001$
PELT	PELTP	0.376	594	317.000	328.500	0.707	1.000	1.000

Table 56: Conover's Post Hoc Comparisons for multiscale segmentation on test sets with conceptual features, z-scores

without conceptual features	with conceptual features	t	df	p
OCP zscores perscale	OCP zscores perscale	-112.793	99	$< .001$
OCP zscores multiscale	OCP zscores multiscale	-113.760	99	$< .001$
OCPB zscores perscale	OCPB zscores perscale	-130.099	99	$< .001$
OCPB zscores multiscale	OCPB zscores multiscale	-116.159	99	$< .001$

Table 57: Comparing segmentations without vs with conceptual features on validation sets, z-scores

without conceptual features	with conceptual features	t	df	p
OCP zscores perscale	OCP zscores perscale	-81.850	99	$< .001$
OCP zscores multiscale	OCP zscores multiscale	-81.718	99	$< .001$
OCPB zscores perscale	OCPB zscores perscale	-86.088	99	$< .001$
OCPB zscores multiscale	OCPB zscores multiscale	-81.711	99	$< .001$

Table 58: Comparing segmentations without vs with conceptual features on test sets, z-scores

segmentation	perscale	multiscale	t	df	p
with conceptual features	OCP	OCP	-3.141	99	0.002
	OCPB	OCPB	0.790	99	0.432
	AUT	AUT	8.185	99	< .001
without conceptual features	OCP	OCP	2.823	99	0.006
	OCPB	OCPB	0.790	99	0.432
	AUT	AUT	.185	99	< .001

Table 59: Comparing perscale and multiscale segmentations, validation sets, z-scores

segmentation	perscale	multiscale	t	df	p
with conceptual features	OCP	OCP	-1.932	99	0.056
	OCPB	OCPB	3.143	99	0.002
	AUT	AUT	8.185	99	< .001
without conceptual features	OCP	OCP	2.361	99	0.020
	OCPB	OCPB	5.901	98	< .001
	7.851	99	< .001		

Table 60: Comparing perscale and multiscale segmentations, test sets, scores

Table 61: Paired Samples T-Test

segmentation	perscale	multiscale	t	df	p
with conceptual features	OCP	OCP	-1.932	99	0.056
	OCPB	OCPB	3.143	99	0.002
	AUT	AUT	8.185	99	< .001
without conceptual features	OCP	OCP	2.361	99	0.020
	OCPB	OCPB	5.901	98	< .001
	7.851	99	< .001		

Table 62: Comparing perscale and multiscale segmentations, test sets, z-scores

		temporal	**scale**
		perscale	**multiscale**
conceptual	yes	OCPB(OCPB)	OCP(OCPB)
features	no	OCP(OCP)	OCP(OCP)

Table 63: Best performing technique for each configuration, assessed on z-scores, in the validation sets (test sets in brackets).

algorithm	temporal scale	segmentation	F1-scores
BEST (bold)	perscale	with conceptual features	0.689 **OCPB**
		without conceptual features	0.177 **OCPB**
	multiscale	with conceptual features	0.686 **OCP**
		without conceptual features	0.175 **OCPB**
AUT	perscale	with conceptual features	0.374
		without conceptual features	0.108
	multiscale	with conceptual features	0.290
		without conceptual features	0.101

Table 64: Best F1-scores and AUT F1-scores for the validation set z-scores

algorithm	temporal scale	segmentation	F1-scores
BEST (bold)	perscale	with conceptual features	0.707 **OCP**
		without conceptual features	0.135 **OCP**
	multiscale	with conceptual features	0.707 **OCP**
		without conceptual features	0.138 **OCP**
AUT	perscale	with conceptual features	0.387
		without conceptual features	0.121
	multiscale	with conceptual features	0.271
		without conceptual features	0.118

Table 65: Best F1-scores and AUT F1-scores for the test set z scores