A COMPUTATIONAL MODEL AND PSYCHOLOGICAL INVESTIGATION OF EVENT SEGMENTATION AND LEARNING

HAMİT BAŞGÖL

BOĞAZİÇİ UNIVERSITY

2021

A COMPUTATIONAL MODEL AND PSYCHOLOGICAL INVESTIGATION OF EVENT SEGMENTATION AND LEARNING

Thesis submitted to the

Institute for Graduate Studies in Social Sciences

in partial fulfillment of the requirements for the degree of

Master of Arts

in

Cognitive Science

by

Hamit Başgöl

Boğaziçi University

2021

A Computational Model and Psychological Investigation

of Event Segmentation and Learning

The thesis of Hamit Başgöl

has been approved by:

Assoc. Prof. Emre Uğur _____
(Thesis Advisor)

Assist. Prof. İnci Ayhan _____
(Thesis Co-Advisor)

Assoc. Prof. Arzucan Özgür _____

Assist. Prof. Esra Mungan _____

Assist. Prof. Nahide Dicle Dövencioğlu _____
(External Member)

June 2021

DECLARATION OF ORIGINALITY

I, Hamit Başgöl, certify that

- I am the sole author of this thesis and that I have fully acknowledged and documented in my thesis all sources of ideas and words, including digital resources, which have been produced or published by another person or institution;
- this thesis contains no material that has been submitted or accepted for a degree or diploma in any other educational institution;
- this is a true copy of the thesis approved by my advisor and thesis committee at Boğaziçi University, including final revisions required by them.

Signature.... .....

Date .....14.06.2021.....

ABSTRACT

A Computational Model and Psychological Investigation

of Event Segmentation and Learning

Event is a fuzzy term that refers to bounded spatio-temporal units, which guide behavior to allow adaptation to complex environments. The study of event segmentation investigates mechanisms behind detecting these spatial-temporal units. Event segmentation theory states that people predict ongoing activities and monitor their prediction errors for segmentation. In this study, the mechanism underlying the event segmentation ability was enlightened with computational models and psychological experiments. Firstly, inspired by event segmentation theory and predictive processing, a computational model of event segmentation and learning was introduced. The performance of the model was compared with humans in point-light displays-based psychological experiments to verify that it can segment ongoing activity into meaningful units, learn them via passive observation, and represent them in its internal representational space. Results indicated that the computational model reached a comparable performance to humans in event segmentation and event representation experiments. Secondly, focusing on the role of prediction errors in event segmentation, several psychological experiments were conducted with the aim of revealing the effect of sensory information (bottom-up processes) and expectation (top-down influence) on perceived event boundaries. Results of psychological experiments were discussed in light of possible implications and future directions.

ÖZET

Olay Ayırma ve Öğrenmenin Hesaplamalı Bir Modeli ve Psikoloji Temelli Bir Araştırması

Olay sınırlı uzaysal ve zamansal bir birime işaret eden belirsiz bir terimdir ve davranışları yönlendirerek karmaşık ortamlara uyumu sağlar. Olay ayırma çalışmaları uzaysal ve zamansal birimlerin tespitinin altında yatan mekanizmaları araştırır. Olay ayırma teorisi ise insanların süregelen aktiviteleri tahmin ettiğini ve olayları birbirinden ayırmak için öngörü hatalarını sürekli olarak izlediğini öne sürer. Bu çalışmada, olay ayırma yeteneğinin altında yatan mekanizma hesaplamalı modeller ve psikoloji deneyleri yoluyla aydınlatılmaya çalışılmıştır. İlk olarak, olay ayırma teorisinden ve öngörülü işlemeden ilham alınarak, olay öğrenme ve ayırmanın hesaplamalı bir modeli geliştirilmiştir. Modelin süregelen aktiviteyi anlamlı bütünlere ayırdığını, onları pasif gözlem ile öğrendiğini ve içsel temsil uzayında temsil ettiğini kanıtlayabilmek için, modelin performansı nokta-ışık görüntülerinden yararlanılarak oluşturulan psikoloji deneyleriyle test edilmiştir. Sonuçlar hesaplamalı modelin olayları ayırma ve temsil etme performansının insanlarla kıyaslanabilir düzeyde olduğunu göstermektedir. İkinci olarak, tahmin hatalarının olay ayırmadaki rolü dikkate alınarak, duyusal bilginin (aşağıdan-yukarıya işleme) ve beklentinin (yukarıdan-aşağıya etki) algılanan olay sınırları üzerindeki etkisi incelenmiştir. Bu kısımdaki araştırmaların sonuçları olası çıkarımlar ve gelecekte yapılacak çalışmalar açısından tartışılmıştır.

ACKNOWLEDGMENTS

First and foremost, I am extremely grateful to my advisors, Dr. Emre Uğur and Dr. İnci Ayhan, for their invaluable advice, continuous support, and endless patience. Their openness to possibilities of science and knowledge in a wide range of areas have encouraged me to complete a study I embrace passionately.

I am also grateful to the Boğaziçi University Research Fund for ensuring the smooth progress of my studies by funding my thesis.

I would like to thank all members of COLORS. Our lab meetings were always inspirational and introduced me to the first steps of building things indistinguishable from magic. I also should thank VisionLab for reminding me of the scope and the complexity of science. I cannot deny the contributions of these groups to my academic progress and, therefore, to the current study.

I also want to thank my friends -Safa, Tuluhan, Ege, and Didem-, who have always been a source of help and receiver of complaints when things would get a bit discouraging and stressful. I especially thank my family for their remarkable tolerance to the never-ending busy work style of the academy.

This study would not be possible without my best friend and soul-mate, Şura. Whenever I need a second mind for a fresh assessment and heart for comprehension, she was there for me.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF ABBREVIATIONS

ADAM: adaptive moment estimation

ANOVA: analysis of variance

CNNs: convolutional neural networks

DTW: dynamic time warping

EST: event segmentation theory

GRNN: gated recurrent neural network

HBT: hierarchy-based technique

MSE: mean squared error

PBT: prediction-based technique

PCA: principal component analysis

PLDs: point-light displays

RELU: rectified linear unit

SOM: self-organizing map

SEM: standard error of the mean

t-SNE: t-distributed stochastic neighbor embedding

CHAPTER 1

INTRODUCTION

Humans experience a continuous stream of information flow and need to chunk their experience into meaningful units to show robust, adaptive, and intelligent behavior. In the spatial experience, they transform scenes into meaningful objects. In the spatio-temporal experience, they divide a sequence of scenes into events, contributing to perception, memory, and action (Zacks, 2020; Zacks & Swallow, 2007). The event, a fuzzy term referring to discrete sequential units that have a beginning and an end, corresponds to a range of diverse and rich structures. For instance, some events are formed around a goal, whereas others do not. Some events are short, whereas other events are large. Events may vary in time from minutes to years (Zacks & Swallow, 2007).

Newtson (1973) conducted initial studies on event segmentation and developed the famous paradigm called unitization. In this paradigm, participants are asked to watch a movie and divide it into meaningful units by pressing a button. Conducting unitization-based psychological studies, Newtson showed that the responses of participants, so-called event boundaries, have a substantial agreement across participants such that the positions of responses are persistent in time. Participants can also detect the smallest (fine-grained) and largest events (coarse-grained) in a movie with and without task instruction, implying that events are hierarchically structured (Hard, Recchia, & Tversky, 2011; Zacks et al., 2001). Even infants are sensitive to event structures. They show increased attention when a goal-oriented event is disrupted (Baldwin, Baird, Saylor, & Clark, 2001).

Subsequent research studying event segmentation with functional magnetic resonance neuroimaging (fMRI) demonstrated that event segmentation is an automatic process in passive observation of activities (Zacks et al., 2001) and during reading (Speer, Zacks, & Reynolds, 2007). That is, observers detect event boundaries automatically. Event boundaries organize episodic memory of observers for objects and scenes (DuBrow & Davachi, 2014; Newtson & Engquist, 1976; Radvansky & Copeland, 2006) and seem to be sufficient to understand and recall the whole event (Schwan & Garsoffky, 2004).

One of the most emphasized topics in the event segmentation literature is why event boundaries have a special place among other frames. Research showed that event boundaries are positions in time that witness a perceptual change, which can be about location, movement, and relative distance or a goal change referring to agents' intentional acts (Cutting, Brunick, & Candan, 2012; Hard, Tversky, & Lang, 2006; Huff, Papenmeier, & Zacks, 2012; Kurby & Zacks, 2008; Newtson, Engquist, & Bois, 1977; Zacks, Speer, Swallow, & Maley, 2010). Perceptual change conveys not only sensory change but also conceptual changes. The interrelation of sensory changes such as simple movement features and conceptual changes such as goals or intentions make event boundaries unique. In other words, event boundaries are determined by the perceptual cycle formed by the bottom-up processing of sensory features and the top-down processing of conceptual knowledge (Neisser, 1976; Zacks, 2020; Zacks & Swallow, 2007).

The second most emphasized topic in the event segmentation literature is the mechanism that results in segmenting events. The most influential theory in the literature is the event segmentation theory (EST) (Zacks, Speer, Swallow, Braver, & Reynolds, 2007; Zacks & Swallow, 2007), which proposes an information processing

model of event segmentation (Zacks et al., 2007) (see Figure 1). According to the EST, event models are working memory representations by which people constantly make perceptual predictions by receiving sensory information as an input. Each prediction generates a prediction error due to the constant uncertainty in the world, and transiently increasing prediction error triggers another event model to predict the following sensory input. Event transition locations are event boundaries; that is, these positions correspond to button responses of participants in the unitization paradigm. Several studies have supported the theory's predictions about the role of prediction errors in segmenting events. In one study, where participants were asked to predict 5 seconds ahead, the authors found that participants made more prediction failure in boundary frames than within-event frames (Zacks, Kurby, Eisenberg, & Haroutunian, 2011). In another study, it was found that predictive eye movements to object in within-event frames are likely to be faster than boundary frames (Eisenberg, Zacks, & Flores, 2018). A recent study using the dwell-time paradigm, which measures the looking time for scenes in a sequence, showed that people spend more time on event boundary frames (Hard et al., 2011). Overall, these studies demonstrate that event boundaries are frames on which people monitor transient prediction error.

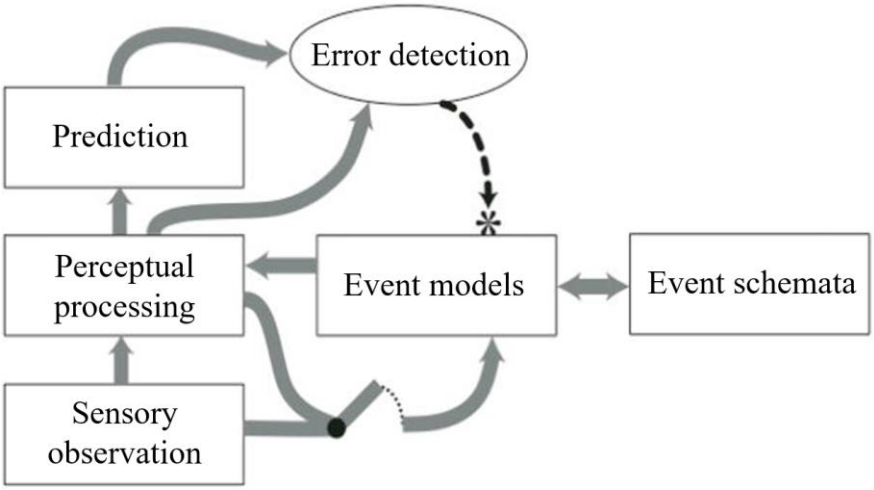


Figure 1 The information processing model of the EST

According to the model, the current sensory information is used to select an event model from the event schemata when the switch is closed. After selecting the new event, the switch is opened again; the new event is used for perceptual processing and perceptual prediction. If the prediction error shows a transient increase, the switch is closed to form a new event. The points at which the switch is closed are called event boundaries (Reynolds, Zacks, & Braver, 2007; Zacks et al., 2007). The figure is adapted from Zacks et al., (2007).

The capability of segmenting events allows humans to use the rich sensory-motor information flow (Richmond & Zacks, 2017). For this reason, studying event segmentation is essential not only for psychology and cognitive science but also for artificial intelligence and robotics (Zacks, 2020). There are several computational cognitive scientific (Franklin, Norman, Ranganath, Zacks, & Gershman, 2020; Gumbsch, Kneissler, & Butz, 2016; Gumbsch, Otte, & Butz, 2017; Metcalf & Leake, 2017; Reynolds et al., 2007), artificial intelligence, and robotics models (Butz, Bilkey, Humaidan, Knott, & Otte, 2019; Gumbsch, Butz, & Martius, 2019; Nery & Ventura, 2011; Wei, Zhao, Zheng, & Zhu, 2016). Despite this interest, a small number of studies compared their models with human performance (Franklin et al., 2020).

In general, this thesis focuses on two topics. One of these is developing a computational model of event segmentation and learning and verifying its

performance by a set of experiments. The second is conducting psychological experiments to explore the perceptual cycle, which refers to the relationship between sensory information and expectation on the perception of event boundaries. Since these two topics lay on similar and different issues, the introduction was prepared to introduce readers to the overlapping concepts.

1.1 Event granularity

People are capable of segmenting events in varying granularities and lengths. Suppose that you are observing a person who is doing the dishes. This activity can be segmented in significantly different ways, such as putting a breakpoint for each cup or completing all activities. While the first segments are called fine-grained events, the second segments are called coarse-grained events. In unitization-based experiments, participants are asked to find the shortest natural/meaningful events for fine-grained events and the largest natural/meaningful events for coarse-grained events (Hard & Tversky, 2003; Newtson, 1973; Zacks, 2004). Coarse- and fine-grained events have characteristic differences. Fine-grained events are more explainable by verbs related to the physical change than coarse-grained events. Conversely, coarse-grained events are more conceivable by verbs implying intentionality (Hard & Tversky, 2003; Hard et al., 2006). When it comes to action segmentation, fine-grained events are described by verbs implying actions, while coarse-grained events are described by nouns implying objects (Kurby & Zacks, 2008). In general, fine-grained events are often involved by coarse-grained events (Hard et al., 2011; Zacks et al., 2001), and both are processed automatically during passive observation (Zacks et al., 2001). According to the EST, humans make perceptual predictions in multiple timescales simultaneously for fine-grained and

coarse-grained event segmentation. As for the theory, the difference between fine-grained and coarse-grained events is their sensitivities to incoming prediction errors. That is, fine-grained event models might be sensitive to minor prediction errors compared to coarse-grained ones, which need more prediction errors for event boundary detection (Zacks et al., 2007).

The perceptual cycle determining event boundaries manifests itself in the context of event granularity. Fine-grained event boundaries primarily depend on the bottom-up processing of sensory features; on the other hand, coarse-grained event boundaries are primarily based on top-down processing of conceptual knowledge (Neisser, 1976; Zacks, 2020; Zacks et al., 2007; Zacks & Swallow, 2007). It is sensible that boundaries between small units (lifting the glass from the table) are readily detectable via sensory change; however, for merging various small segments in a larger event segment, conceptual knowledge might be required (lifting the glass from the table and going out from kitchen). While fine-grained segmentation boundaries are detectable by tracking movement features, coarse-grained segmentation boundaries are discernible by tracking goals and intentionality related to conceptual change. In the present studies, one or two sides of this perceptual cycle were manipulated to understand the relative relationship between the bottom-up and the top-down processes in detecting event boundaries.

1.2 Sensory information and conceptual knowledge

What is the mechanism behind the relationship between the two sides of the perceptual cycle? Research suggests that top-down influence makes segmented events coarser. For example, predictable behaviors are generally segmented into coarser units than unpredictable ones (Newtson, 1973; Wilder, 1978b, 1978a).

Participants segment events into coarse-grained units when having prior knowledge (Graziano, Moore, & Collins, 1988; Markus, Smith, & Moreland, 1985). In line with these studies, researchers investigated the relationship between movement features and their interpretation using simple shapes. For instance, Hard et al. (2003; 2006) studied ambiguous event segmentation by using chase and hide-and-seek movies depicting social interactions by simple movements (see Figure 2) (Heider & Simmel, 1944). They manipulated the knowledge of people on depicting events by controlling the number of observations. In their first observation, people relied on physical changes and unitized finer events, whereas, in the fifth observation, they detected coarser events, relying on conceptual knowledge. The study showed that learning (top-down influence) changed the way people interpret movies and therefore perceived event boundaries (Hard & Tversky, 2003; Hard et al., 2006).

Additional evidence of the role of top-down influence on perceived event boundaries comes from the study conducted by Zacks (2004), who investigated the effect of intention on the segmentation of simple movements of a circle and rectangle (see Figure 2). In his first experiment, Zacks (2004) found that people who think that movements of shapes were generated by an equation observed more event boundaries than people who think that shapes were managed by players. Although the subsequent experiment could not find this effect, it remained clear that perceiving intentionality changed the way of processing movement features in the detection of event boundaries (Zacks, 2004). In this study, participants seeing intentional movements segmented events by giving prominence to the relative distance of shapes; on the other hand, those seeing movements generated by an equation focused on the acceleration of a particular shape (Zacks, 2004).

These studies (Hard & Tversky, 2003; Hard et al., 2006; Zacks, 2004) concentrated on how top-down processes affect the perception of event boundaries, namely, on the one side of the perceptual cycle. The other side of the perceptual cycle considers the role of sensory information in generating perception. In this thesis, computational as well as psychological experiments were conducted with a focus on examining the effect of sensory reliability on perceived event boundaries.

Predictive processing and prediction error signals present a meaningful interpretation of seminal studies (Graziano et al., 1988; Hard & Tversky, 2003; Hard et al., 2006; Markus et al., 1985; Newtson, 1973; Wilder, 1978b, 1978a; Zacks, 2004) and illuminate the sensory part of the perceptual cycle.

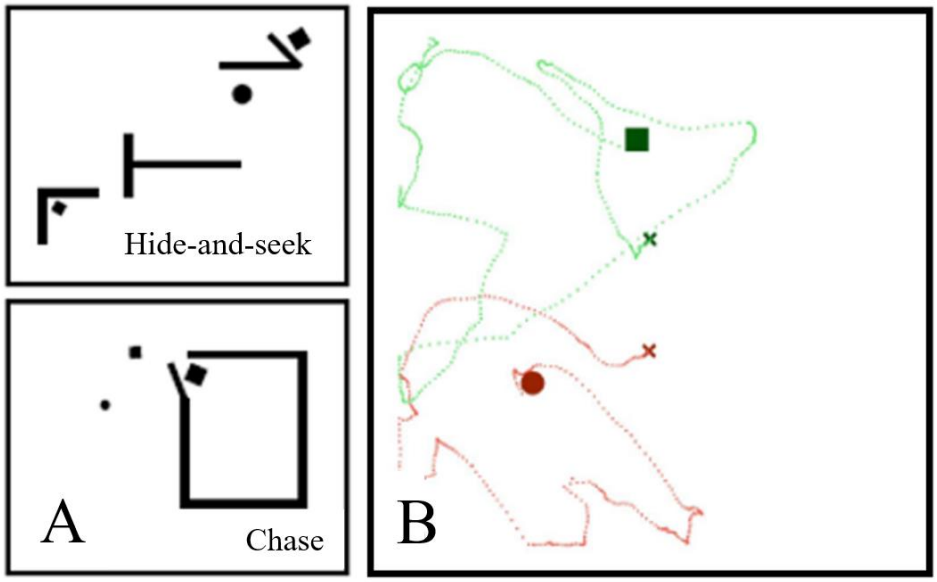


Figure 2 Simple stimuli used in reference papers

(A) shows interactions of simple shapes in Hard et al. (2003; 2006). From interactions of shapes, one can perceive social interaction. (B) displays a sample of movements of two simple shapes used in Zacks (2004). The figure is adapted from Hard et al. (2003; 2006) and Zacks (2004).

1.2.1 Predictive processing and prediction error signals

Predictive processing is a unified framework put forward to explain cognition as a whole. As for the framework, the brain generates mental models to predict the current sensory information and updates mental models via prediction error signals (De Lange, Heilbron, & Kok, 2018; O'Reilly, 2013; Wiese & Metzinger, 2017). From the perspective of predictive processing, event models proposed by the EST are similar to mental models aiming to predict the current sensory input.

Predictive processing assumes that perception is a construction process that is emerged from the interaction between bottom-up and top-down processes. While bottom-up processing of sensory information provides necessary ingredients of perceptual content, top-down processing of influence integrates those depending on the prior knowledge, which can be modeled in a Bayesian manner (O'Reilly, 2013; Wiese & Metzinger, 2017). The brain is a prediction machine that attempts to guess incoming sensory inputs with top-down expectations (Clark, 2013), but how expectations influence perception? When sensory information is noisy and weak, expectations manage and modulate the perception of agents. For instance, a cloud of dots' perceived direction can be changed according to the expectation if the incoming sensory information is noisy (Chalk, Seitz, & Series, 2010; Kok, Brouwer, Van Gerven, & De Lange, 2013). In such cases, the top-down influence dominates perception. Conversely, if the incoming information is evident or the current expectation is weak and unreliable, the expectation resulting from top-down influence cannot dominate perception (De Lange et al., 2018; O'Reilly, 2013). In short, the contribution of expectation and sensory information to perceptual construction depends on their relative reliabilities (confidences). The influence of prior knowledge on perception is at most when expectations are reliable, but stimuli are ambiguous; on the other hand, its effect is lowest when expectations are weak and stimuli are reliable. From the

perspective of predictive processing, this harmony is achieved by the dynamic regulation of prediction error signals. When sensory information is unreliable, the valence of the prediction errors is reduced by giving rise to the high valence of expectations. On the contrary, when it is reliable, the valence of prediction errors increases, which results in the high valence of sensory information (Bastos M. et al., 2012; De Lange et al., 2018; Feldman & Friston, 2010). This point of view can make the results of already mentioned studies clearer in the light of event segmentation (Graziano et al., 1988; Hard & Tversky, 2003; Hard et al., 2006; Markus et al., 1985; Newtson, 1973; Wilder, 1978b, 1978a; Zacks, 2004).

1.2.2 Prediction error signals and event boundaries

According to the EST, event models are working memory representations by which one can make predictions and generate expectations. While predicting with the help of event models at hand, the system monitors prediction error signals coming from these predictions for revealing event boundaries. Employing the reasoning borrowed from predictive processing, if the sensory information is not reliable, people should rely on their expectations by reducing the importance of prediction errors, which should cause coarser event models and, therefore, coarser event segments. In brief, unreliable sensory information should lead people to form coarser event segments if they think their expectations are reliable. Indeed, the results (Graziano et al., 1988; Hard & Tversky, 2003; Hard et al., 2006; Markus et al., 1985; Newtson, 1973; Wilder, 1978b, 1978a; Zacks, 2004) pointed out this direction. Despite the apparent role of prediction errors regulating the perceptual cycle, to the best of our knowledge, there is no study aiming to inspect the effect of the *relative* reliability of sensory inputs and expectations on perceived event boundaries.

Experiments designed for investigating the two sides of the perceptual cycle require an easily manipulable and informationally rich stimulus. For this reason, point-light displays were used for both psychological and computational experiments in order to express meaningful sequential information of natural human behaviors.

1.3 Biological motion perception and point-light displays (PLDs)

Animals have a strong tendency towards biological movements such as motor actions, locomotion, facial expressions, and gestures because perceiving them has a substantial survival value (Giese & Poggio, 2003; Johansson, 1973; Troje & Basbaum, 2008). Probably, one of the most compelling pieces of evidence for this claim is that humans can perceive biological motion even from moving dots. Johansson (1973) invented the famous point-light walker, depicting human walking by several points without pictorial information. With the help of relative movement of points, people perceive various kinds of human movements, emotions, actions (Alaerts, Nackaerts, Meyns, Swinnen, & Wenderoth, 2011), and gender (Troje & Basbaum, 2008).

Since PLDs can convey meaningful sequence information by a set of points, they are easily manipulable for psychological experiments. Besides their benefit, they are noticeably advantageous for developing computational models because they reduce data dimensionality and processing time. PLDs have a history in computational modeling of event segmentation (Metcalf & Leake, 2017; Reynolds et al., 2007) and empirical research (Hemeren & Thill, 2011; Noble et al., 2014). In computational and psychological experiments, PLDs were used to represent human behaviors.

1.4 The aim of the thesis

So far, the general content closely related to two studies in this thesis has been described. In general, the main aim of this thesis was to investigate the perceptual cycle from the perspective of predictive processing. In the first part of this thesis, a computational model of event segmentation and learning was developed, and its event segmentation performances were compared with humans. In this part of the study, only one side of the perceptual cycle was manipulated, namely sensory information. It was manipulated to reveal (1) whether a reduction in sensory reliability would make segmented events coarser and (2) whether a similar effect would be seen in the decisions of the computational model. As further validation, event representations of the computational model were compared with those of humans.

In the second part of the thesis, the relative effect of sensory information and expectation on perceived event boundaries was investigated by psychological experiments. Two sides of the perceptual cycle were manipulated to reveal their interaction in the perception of event boundaries. In general, it was expected that the reliable expectation would decrease the valence of prediction error, and reliable sensory input would upgrade the importance of prediction error. In particular, it was predicted that reliable expectation along with unreliable sensory input would reduce, whereas unreliable expectation would increase the number of perceived event boundaries.

CHAPTER 2

A COMPUTATIONAL MODEL OF EVENT SEGMENTATION AND LEARNING

The EST proposes an information processing model of event segmentation learning to explain how people segment continuous information flow into discrete events (Reynolds et al., 2007; Zacks, 2020; Zacks et al., 2007). According to the EST, event models are working memory representations by which people constantly make perceptual predictions by receiving sensory information as an input. Each prediction generates more or less a prediction error due to the constant uncertainty in the world. The transiently increasing prediction error marking the current event model being incompetent for making predictions triggers another event model to predict the following sensory input. The event model transition locations determine event boundaries in time and correspond to button responses of participants in the unitization paradigm.

In line with the EST, predictive processing appeals to prediction error. Predictive processing is a conceptually unified framework aiming to explain cognition as a whole. It asserts that the brain generates mental models to predict current sensory information and develops mental models by learning from prediction error signals (Clark, 2013; O'Reilly, 2013; Wiese & Metzinger, 2017). The framework gives central importance to the top-down influence of mental models generated by the brain. Event models proposed by the EST are comparable to mental models introduced by predictive processing, and they share certain similarities. Both event models and mental models make a prediction, contribute to sensory information, and represent knowledge. Recent studies search for collaboration

between event cognition and predictive processing (Gumbsch et al., 2017; Hohwy, Hebblewhite, & Drummond, 2021; Stawarczyk, Bezdek, & Zacks, 2021).

Several computational models of event segmentation were developed in cognitive science (Franklin et al., 2020; Gumbsch et al., 2019, 2016, 2017; Metcalf & Leake, 2017; Reynolds et al., 2007). For instance, Reynolds et al. (2007) developed a set of computational models in order to investigate the assumptions of the EST and tested those models by PLDs. Their primary model was based on a gated-recurrent neural network (GRNN), adjusting its internal representation with the help of a gating mechanism. In one of their simulations (IIIB), the gating mechanism was signaled by either a supervised (using ground-truth event change) or self-supervised (using an equation over prediction error signal statistics) manner for the event change. Even though both models demonstrated that the EST proposes a plausible information-processing model of event segmentation, they could not learn and segment human behaviors in varying granularities. Moreover, transitions between human behaviors used by Reynolds et al. (2007) included unnatural interpolations and abruptions dissimilar to the continuous information stream people experience. Improving the model proposed by Reynolds et al. (2007), Metcalf and Leake (2017) added a reinforcement learning agent controlling the gating mechanism of GRNN. The involvement of the reinforcement learning agent made the equation used by Reynolds et al. (2007) unnecessary. Even though Metcalf and Leake (2017) showed that their model outperformed the classical model developed by Reynolds et al. (2007), it faces similar challenges.

Gumbsch et al. (2016, 2017) developed a model of event segmentation and learning to chunk sensory-motor information flow for learning and planning in robots. They represented events by a set of linear models corresponding to different

sensory dimensions, which hindered the model from representing events in a multimodal manner and discovering the possible relationships between different sensory modalities in an event unit. Being designed to interact with the world, these models were not developed for passive observation, even though the literature on event segmentation is primarily dependent on this observation type (Hard & Tversky, 2003; Hard et al., 2006; Newtson, 1973; Newtson & Engquist, 1976; Newtson et al., 1977; Zacks, 2020; Zacks et al., 2001).

To the best of my knowledge, there is no event segmentation model whose segmentation decisions are directly compared with human data except the one developed by Franklin et al. (2020). Even though this is not necessarily a limitation for robotic models interacting with the world (Gumbsch et al., 2019, 2016, 2017), it is essential for computational models as an assessment criterion (Metcalf & Leake, 2017; Reynolds et al., 2007). Franklin et al. (2020) utilized a neuro-symbolic neural network approach to capture human event cognition and its domains such as event memory, segmentation, retrieval, and inference. Segmentation decisions of the model were compared to human data in naturalistic videos, distinguishing the model from other models discussed. Although they showed that their model could capture a wide range of phenomena in event cognition, including event segmentation, it had certain limitations. Firstly, the correlation score of the computational model with human data was not satisfactory; secondly, segmentation decisions of the model for varying granularities were not compared to human data. Finally, the hierarchical nature of event segmentation was beyond its capabilities.

Inspired by models developed by Gumbsch et al. (2016, 2017), a computational model of event segmentation and learning was developed in this study. The proposed computational model consists of multi-layer perceptrons (i.e.,

feed-forward neural networks having more than one layer) representing and learning to predict event segments. These multi-layer perceptrons, called event models, are regulated by a computational architecture. Bearing a resemblance to the suggested cognitive system by the EST, the computational architecture tracks prediction errors generated by event models for event model transition and, consequently, determining event boundaries. The computational architecture also decides when a new event model is generated, and old event models are trained. In line with predictive processing, event models are trained to reduce prediction error signals to represent the world. Their aim, reducing incoming prediction error signals, connects the proposed computational model with the predictive processing framework. The proposed model is fully self-supervised and learns the relationship in the data by utilizing the prediction error signals, meaning that it does not need explicit human-made labels for learning and segmenting events.

As our contribution, the proposed model can produce multimodal event segments in varying hierarchies via passive observation, learn those segments in a self-supervised manner with the help of multi-layer perceptrons, and represent these events in a meaningful way in its latent dimensions. Besides computational experiments, it was shown by two online PLDs-based psychological experiments on event segmentation decisions and event similarity judgments that performances of the model were comparable to those of participants. Also, those online psychological experiments were used to assess whether (1) the reduced rate of change and sensory reliability would decrease the number of perceived event boundaries and (2) similarity is a meaningful relationship in the context of events. Specifically, in the psychological experiments of event segmentation, the unitization paradigm (Newtson, 1973) was used to receive event segmentation decisions. In the

psychological experiments of event representation, the pairwise ranking paradigm was conducted for gathering human similarity judgments between events.

In computational and psychological experiments of event segmentation, roles of sensory reliability and change were examined. In the literature, it was suggested that sensory reliability reduces the effect of prediction error (De Lange et al., 2018) from the perspective of the EST, and therefore, decreases the number of perceived event boundaries. It was also shown that change increases the likelihood of perceiving event boundaries (Cutting et al., 2012; Hard et al., 2006; Huff et al., 2012; Newtson et al., 1977; Zacks et al., 2010), and change is greatest at event boundaries (Hard et al., 2011). To test this, two types of videos, namely normal and noisy, were generated with the purpose of investigating the effect of sensory reliability on perceived event boundaries by psychological experiments. The normal video included PLDs expressing rich human behaviors such as jumping, push-upping, and complex object searching. In contrast, the noisy video was generated by a smoothing operation applied on normal video, which reduces the change between successive timesteps. In unitization-based psychological experiments, participants' segmentation decisions for normal and noisy videos for different granularities (i.e., fine and coarse-grained segmentation). In general, in this part of the study, it was analyzed that whether event boundary decisions of the model (1) are comparable to those of participants in terms of number and locations of boundaries, (2) are affected by the video similar to participants' event boundary decisions, and (3) ambiguous sensory information makes perceived event segments coarser. Results of computational experiments on event segmentation showed that, despite occasional mismatches, the proposed computational model performed well in segmenting human behaviors in varying granularities and learning them. Moreover, ground truth from psychological

experiments demonstrated that segmentation decisions of the model are mainly consistent with humans.

In the psychological experiments of event representations, an online experiment was conducted for receiving pairwise similarity judgments from people as ground-truth data. Then, these results were analyzed to find out (1) whether the notion of similarity is meaningful in the context of event cognition and (2) whether the proposed mode captured event similarity judgments of people in its low dimensional latent representations. Results concluded that event similarity judgments of participants showed a remarkably high correlation, and the proposed model also successfully captured semantic relationships between events.

2.1 Computational model

In this part, how the proposed computational model segments a sequence of information into different events (event segmentation) and how it organizes the relationship between events in its latent dimensions (event representations) were analyzed. Furthermore, its capabilities were assessed by the ground-truth data received from psychological experiments.

2.1.1 Proposed model

The proposed model was based on the computational architecture proposed by Gumbsch et al. (2016, 2017). Researchers developed a computational architecture consisting of event models making predictions for the following sensory information and a generic system regulating these event models by considering their respective prediction error signals. Each event model involves a set of linear models corresponding to the sensory modalities received as input, and each linear model

predicts one type of sensory modality. The prediction error signal is utilized for linear model training and calculating a dynamically changing threshold value, called surprise threshold. When the prediction error of an active linear model is higher than its surprise threshold (i.e., the current forward model is unqualified for current sensory prediction), the system enters into the search period to select among different linear models and continues to predict.

Consistent with the suggestion of the EST, the architecture controls granularities of events to be segmented by changing the sensitivity of models to incoming prediction error signals. According to the EST, fine-grained events are highly prone to incoming prediction error, whereas coarse-grained events are insensitive. Similar to the EST, the surprise threshold formula includes a parameter called the confidence parameter to regulate the sensitivity of models to prediction error signals. The confidence parameter regulates the tolerance of models to the received prediction error. Lower confidence parameters lead the system to enter the search period more frequently and, consequently, detect more event boundaries and event segments (Gumbsch et al., 2017).

In their model (Gumbsch et al., 2016, 2017), each event model consists of linear models responsible for predicting one modality; in particular, motor modality plays a role in event formation. These properties bring several challenges to their model to capture the performances of humans in segmenting activities via passive observation. Firstly, events are complex structures that are composed of associations between different modalities. For instance, visual and auditory information goes hand-in-hand in our everyday experience, but their model cannot capture associations. Secondly, the relationship between time and a sensory dimension is not always linear. For example, even in their simplest form, human behaviors can be

expressed only by nonlinear functions. Finally, the event segmentation literature is based on passive observation (Newtson, 1973), which does not necessitate motor modality activation during event segmentation.

A general overview of the proposed model and an example of stimuli used in this study were given in Figure 3. In general, (1) event models were changed with a multi-layer perceptron capable of approximating nonlinear functions. In this way, the proposed model was altered to approximate complex sequences and learn associations between possible modalities. Also, (2) the role of motor modality was reformulated to segment events during passive observation.

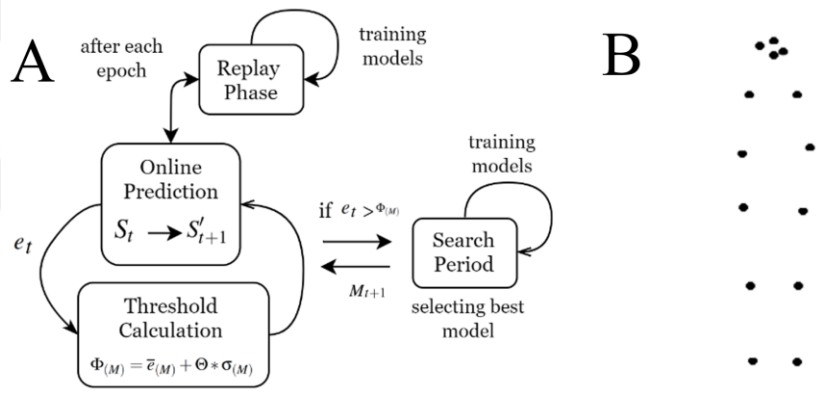


Figure 3 The overview of the computational model and PLDs

(A) The overview of the proposed model. In the online prediction phase, the current event model makes a prediction, the current prediction error is calculated, and the surprise threshold is computed. If the current prediction error is more than the surprise threshold, the system enters the search period where the best model is returned for online prediction. (B) shows a point-light display representing a human figure.

2.1.1.1 Online prediction

The proposed architecture has one active event model M_t at time t. The event model M_t predicts the change $\Delta S'_{t+1}$ observed in the sensory input. Therefore, the predicted sensory observation is computed by (1).

$$S'_{t+1} = S_t + \Delta S'_{t+1} \quad (1)$$

Given the observations, the active model M_t makes online predictions by sensory input, learns from prediction error signals by changing neural network weights (i.e., multi-layer perceptron) via backpropagation, stores them, and calculates a dynamically changing threshold with stored errors. If the error of M_t exceeds the surprise threshold, which is a sign of incompetency, the computational model enters the search period to return a competent event model for predicting the following sensory input.

2.1.1.2 Surprise threshold and search period

Similar to Gumbsch (2016, 2017), for each generated event model, our model has a surprise threshold φ_M determining how much prediction error can be tolerated, and when an event model is not suitable for predicting the following sensory observation. The threshold φ_M is calculated by the rolling mean of stored prediction errors $\bar{e}_M$ and of the variance $\overline{\sigma}_M$. The rolling mean has a window w. The confidence parameter θ regulates the coarseness of the event to be segmented. φ_M is calculated by (2).

$$\varphi_M = \bar{e}_M + \theta * \bar{\sigma}(\bar{e}_M) \quad (2)$$

If the prediction error of an event model exceeds φ_M, the system enters into the search period for the purpose of finding a suitable event model M_i. At the beginning of the searching period, a potential event model is generated with random weights. All event models on hand are then trained for rehearsal duration that amounts to the number of epochs used in the search period phase. After the training procedure is over, the event model receiving the least mean squared error value is

returned for online prediction, the new event model is removed if it is not the best one. The effect of training on all event models except the best one is removed.

In order to return the best possible event model from the search period, the training set should include the next n timesteps in the sequence, that is $S_{t:t+n}$. Simultaneously, event models should not forget their sensory experience history, which can be referred as M_{is}. This is particularly challenging as multi-layer perceptrons are prone to catastrophic forgetting (French, 1999), which defines a situation in which novel information erases the information learned. For this reason, a training set for each event model was sampled from the combination of $S_{t:t+n}$ and M_{is}.

2.1.1.3 Memory range and replay

When encountered with a surprise signal, the system enters a search period. The extensive search in the search period to return the best event model increases the training time exponentially as a function of event models in the system. To overcome this problem, following each training epoch, the system detects event models not used for n epochs (i.e., memory range) and removes them to improve computational efficiency.

Another problem the system faces is that the online training regimes of multi-layer perceptrons are not stable. To foster memory consolidation, avoid catastrophic forgetting, and reduce training time, a replay phase was introduced, in which each event model M_i used in the recent epoch are trained by M_{is}. A replay phase was observed in the hippocampal regions of the brain, and it is mainly used for memory consolidation (Ólafsdóttir, Bush, & Barry, 2018). Besides its role in the brain, it was

suggested as a computational technique to stabilize the training process of reinforcement learning agents (Andrychowicz et al., 2018).

So far, the formalization of the computational model has been described; next, the sequential information given to the computational model and shown to the participants will be detailed.

2.1.2 Dataset preparation with PLDs

Natural human behaviors in computational and psychological experiments were taken from the KIT Motion-Language Dataset (Plappert, Mandery, & Asfour, 2016). The overall activity included 12 human behaviors, such as walking, jumping, picking an object, sitting on a chair, and searching for an object. Behaviors were determined to be rich as much as possible to determine the genuine capabilities of the model with respect to ground truth received from human data. Selected behaviors were represented in PLDs format using X and Y dimensions of 14 markers (RFHD, LFHD, LBHD, RBHD, RSHO, LSHO, RELB, LELB, RWRA, LWRA, RKNE, LKNE, RTOE, LTOE). Selected behaviors were added back-to-back through interpolating the marker positions. Here, to prevent participants and the computational model from exploiting unnatural interpolation trajectories as segmentation cues, two behaviors were added back-to-back if the distance between the end of the first and the start of the second behavior is the smallest option in possible permutations. One exception to this permutation was the control behavior selected for assessing data reliability. The control behavior appeared in a video two times. Its first appearance was between 42.75-71.5 seconds, and its second appearance was in 209.25-237.5 seconds. The control behavior can be defined by “lifting and lowering an object by hand”; it was selected because it involves sharp

hand movements in the Y dimension, which serves precise sensory information for event segmentation decisions.

This was the preparation of the normal dataset. Another dataset, namely a noisy dataset, was created to investigate the role of change and sensory reliability in event boundary detection in humans and the proposed computational model. The preparation of the noisy dataset was explained after its purpose was clarified in the following sections.

2.2 Event segmentation experiments

The proposed computational model was tested both for event segmentation and event similarity judgments. In the following subsections, psychological and computational experiments on event segmentation were described.

2.2.1 Psychological experiments of event segmentation

In psychological experiments of event segmentation, two experimental conditions were determined. One was event granularity (fine-grained and coarse-grained segmentation), and the other was sensory reliability (normal input, noisy input). The dependent variables were the number and positions of the event boundaries.

2.2.1.1 Event granularity

People are capable of segmenting into events with varying granularities following the task instructions in psychological experiments. For fine-grained events, participants are asked to find the shortest natural/meaningful events. On the other hand, for coarse-grained events, they are asked to find the largest natural/meaningful events (Newtson, 1973; Zacks & Swallow, 2007).

2.2.1.2 Event segmentation, change, and reliability of sensory information

There is an intrinsic relationship between the change and event segmentation (Cutting et al., 2012; Hard et al., 2011, 2006; Huff, Meitz, & Papenmeier, 2014; Kurby & Zacks, 2008; Newtson et al., 1977; Zacks, 2020; Zacks et al., 2010; Zacks & Swallow, 2007). Specifically, Hard et al. (2011) demonstrated that changes at event boundaries are more numerous than those at other frames. At the same time, changes are maximal at the event boundaries of coarse-grained units.

From the perspective of predictive coding, the relative reliability of expectations and sensory information determines perception. Ambiguous sensory information (i.e., noisy video) reduces the effect of perceived prediction error (De Lange et al., 2018). Therefore, from the perspective of the EST, it should decrease the number of perceived event boundaries. That is, reduction in change should also decrease the number of perceived event boundaries. With the aim of verifying (1) this information in the literature and (2) assessing whether the proposed model's event boundary decisions are affected similarly to participants' decisions, two datasets/videos -normal and noisy- involving the same behaviors were prepared.

While the normal dataset/video was prepared by directions mentioned in the dataset preparation part, the noisy video was prepared by reducing the change in the normal video by Gaussian noise (window: 40, standard deviation: 10). This resulted in a video whose fine dynamics were smoothed, and movements of points were perceived as much more fluent. In this way, two 16 Hz and 270-second videos were created. The experiment was prepared in Psychopy3 (Peirce et al., 2019) and conducted on an online platform named Pavlovia.

2.2.1.3 Participants

The experiment had two experiment conditions: sensory reliability (normal input, noisy input) and event granularity (fine-grained, coarse-grained segmentation). Nineteen participants (9 female, mean age 25) were recruited for a within-subject design. Being selected via convenient sampling voluntarily, participants were primarily undergraduate and graduate university students. None of the participants had a problem with vision. Participants received a chance to be eligible for the 150 Turkish liras Amazon gift voucher lottery. Participants' anonymity and confidentiality were established by separating any possible personal references from required experimental data. Since participants were native speakers of Turkish, all experimental materials from informed consent to experimental instructions were prepared in Turkish. The study was consistent with the requirements of research ethics and confirmed by the Boğaziçi University Ethics Coordinating Committee (see Appendix A).

2.2.1.4 Procedure

Participants were reached from the internet, informed about the experiment to a short extent, and asked whether they were voluntary for participating in the study. Upon their agreement, they were sent a link generated by the online experiment platform (i.e., Pavlovia). They could read and accept informed consent, fill in a demographic form and a short questionnaire, and complete the experiment from the provided link. In the informed consent, they were informed of the voluntary nature of the experiment and allowed to leave the experiment in case they felt uncomfortable. The short questionnaire included questions about (i) their demographic background, whether (ii) they use glasses or lens, (iii) they have any psychological or neurological

problems and (iv) use any drugs for treatment. Then, participants were informed about the experiment, such as what event segmentation is, and experimental instructions, such as the requirement of attending the experiment in a quiet and proper place.

Each participant firstly segmented the normal video and then the noisy video; for the segmentation of two videos, the experiment was the same. The experiment started with the instruction of the segmentation granularity condition, asking participants to segment the video into either the shortest, natural, and meaningful or the longest, natural, and meaningful events by pressing the space button. The level of segmentation granularity was counterbalanced. Participants were observed and segmented the video two times for each level of granularity. In the second observation, participants were asked to segment events in the shortest or longest possible way according to the granularity level to receive sound ground-truth data for the computational model. Throughout the study, these observations were coded for maintaining simplicity, such as Fine 1, Fine 2, Coarse 1, and Coarse 2. For instance, Fine 1 refers to the first observation of the fine-grained segmentation level.

Participants were informed that the system coded their decisions by the appearance of a grey rectangle at the bottom of the video due to pressing the space button. A captured scene that was shown to the participants was given in Figure 4.

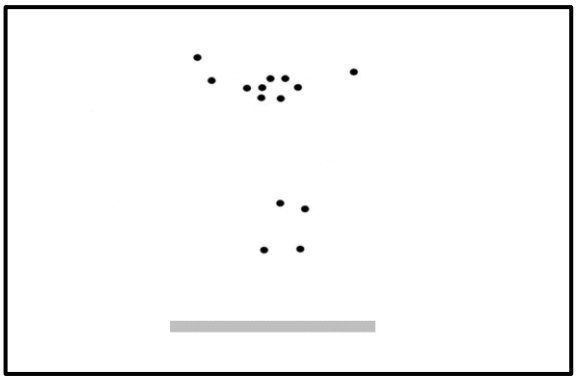


Figure 4 Screen of participants in event segmentation experiment

Each participant attended the experiment by using the online experiment link provided. After receiving the required instructions, they were shown a movie composed of moving dots whose movements were changed according to the level of sensory reliability (noisy or normal). The depicted human figure is looking slightly upwards with its arms raised. When the participant pressed the space button throughout the experiment, a grey rectangle appeared, informing the participant that the response was received.

Subsequent to the segmentation of movies in a level of segmentation granularity, participants were asked to answer a two-choice question asking, “Which way did you segment the video you have seen?” with two possible answers “the shortest, meaningful and natural way” and “the longest, meaningful and natural way,” to detect participants who could not attend the experiment. After then, they received open questions such as (i), please describe what you have seen from the videos shown during the experiment, (ii) have you ever seen similar images to those you were presented in the videos, and (iii) please mention your comments and impressions about the experiment. Following open questions, they ranked two expressions in a subjective rating scale (1-5) in accord to their perception: (i) I thought that movements of points express a human figure, and (ii) I thought that movements of points express certain human behaviors (walking, eating, or jumping, etc.). Finally, participants were requested to write their emails to be considered for the Amazon gift voucher lottery.

2.2.1.5 Results

Online psychological experiments pose additional challenges for psychological experiments. One of those challenges is data reliability (Gosling & Mason, 2015). For this reason, several control measures were employed before applying statistical tests on the data to maintain data reliability.

Firstly, whether participants relied on the segmentation granularity instruction was checked by comparing their responses in Fine 2 and Coarse 2 levels. According to the instruction, they should produce more responses in Fine 2. For this reason, two participants were excluded from the analysis as their number of responses were more in Coarse 2 than Fine 2 in normal video segmentation. On the other hand, examining the same for noisy video segmentation did not result in any reliability problem. Secondly, whether the responses of the remaining participants were reliable in fine-grained and coarse-grained segmentation levels was checked by comparing them. Participants' responses were turned into a continuous representation format by a mixture of Gaussian distributions, each corresponding to one button response, to compare two discrete sets of responses. For fine-grained segmentation, the normal distribution generated for each button pressing was determined to be $N(t, 1)$, whereas, for coarse-grained segmentation, it was determined to be $N(t, 4)$ as the transition between two coarse-grained units might take more time than that of two fine-grained units. For exploring the reliability in coarse-grained segmentation, Gaussian response distributions of Coarse 1 and Coarse 2 were compared by Pearson correlation coefficient (r). Responses of participants showing very weak or inverse ($r < .1$), but significant correlation ($a = .05$) were excluded. This procedure resulted in the exclusion of three participants. The responses of the remaining participants were tested for two parts of control behavior in Fine 2 by employing the same method.

This operation, along with outlier checking ($z > 2.98$), did not result in data exclusion. There remained 14 participants whose event segmentation decisions were considered for hypothesis testing.

The assumption of normality was checked for each possible group (Event Granularity x Sensory Reliability x Observation Order) by Shapiro–Wilk test, which showed that response distributions of Fine 1 ($W = 0.846$, $p = .019$) and Fine 2 ($W = 0.798$, $p = .005$) in the segmentation of normal video were significantly non-normal. On the other hand, Fine 1 ($W = 0.855$, $p = .026$) was non-normal, too. Considering this fact, Friedman's ANOVA was used in order to detect any statistically significant difference between groups. Analysis revealed that the number of responses significantly and meaningfully differed across groups ($X^2(7) = 80.93$, $p = .000$).

Aiming at finding out specific group differences, Wilcoxon signed-rank test was applied, at the same time, using Holm–Bonferroni method to adjust p values in order to avoid from inflating Type 1 error rate. One of the hypotheses, the number of event boundaries produced by participants would be less in the noisy video than the normal video, turned out to be true, as shown in Figure 5. Participants perceived a smaller number of event boundaries in noisy video ($M = 32.71$, $Mdn = 17.5$, $SEM = 4.44$) than in normal video ($M = 41.91$, $Mdn = 22.5$, $SEM = 5.95$), $W = 879.0$, $p = .021$, $r = .191$. Similarly, an identical trend was seen in the fine-grained and coarse-grained segmentation of videos. For the fine-grained segmentation level, the normal video ($M = 68.14$, $Mdn = 43$, $SEM = 9.56$) received more response than the noisy video ($M = 52.50$, $Mdn = 43.5$, $SEM = 6.98$), $W = 288.5$, $p = .016$, $r = .31$. Similar to fine-grained, the coarse-grained segmentation level saw the same trend, normal ($M = 15.67$, $Mdn = 16$, $SEM = 1.49$) and noisy video ($M = 12.92$, $Mdn = 10.5$, $SEM = 1.66$), $W = 203.5$, $p = .018$, r = .33, with a higher effect size. Results verified that

change and reliability of sensory information affect the number of perceived event boundaries. Effect size calculations (*r*) for Wilcoxon signed-rank tests were based on Rosenthal (2011, p. 19).

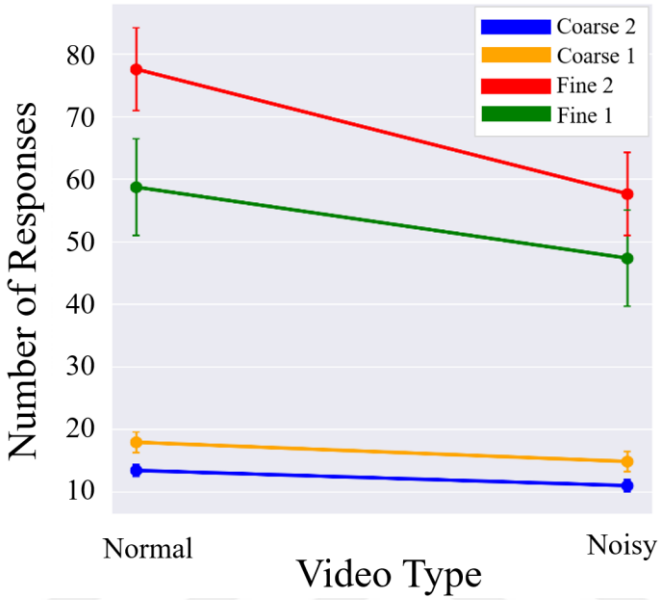


Figure 5 Overall results of the experiment of event segmentation

The mean number of responses of each group with corresponding standard errors of the means were given (+/- 2 *SEM*s). *SEM*s were normalized by the method proposed by Cousineau (2005) to remove the effect of individual variability on error bars.

As for the observation order, observing normal and noisy videos two times in coarse-grained segmentation condition reduced the number of responses (Coarse 1: $M = 16.39$, $Mdn = 14.5$, $SEM = 1.70$ and Coarse 2: $M = 12.21$, $Mdn = 11.5$, $SEM = 1.39$, $W = 351.0$, $p = .000$, $r = .59$), as expected. In contrast, in fine-grained segmentation condition, the number of responses increased in the second observation (Fine 1: $M = 53.03$, $Mdn = 38.5$, $SEM = 7.57$ and Fine 2: $M = 67.60$, $Mdn = 47.0$, $SEM = 9.14$, $W = 25.0$, $p = .000$, $r = .51$).

Overall, results demonstrated that the reduced change removing fine dynamics of behaviors and unreliable sensory information decreased perceived event

boundaries. The second observations were considered throughout the result section since responses of participants in the second time would be more reliable as ground-truth data than their first observations.

The results showed that disrupting and smoothing fine dynamics of observed actions resulted in detecting coarser event segments. Another question might be about the role of change in detecting event boundaries. Hard et al. (2011) suggested that the degree of change was maximal at coarse-grained event boundaries. With the aim of assessing this finding, event boundary histograms representing response probabilities of participants over time were computed. Event boundary histograms were generated by grouping responses of participants into *n*-second bins and normalizing each bin value by the number of participants. The unitization paradigm requires participants to report observed event transitions by a button. The reporting process takes time because of the decision-making process and/or the activation of motor modality. For this reason, correlations were computed by shifting response distributions backward in time to assess the relationship between change and response probabilities of participants in-depth. Response probabilities were revealed by computing event histograms with 0.25-second bins. Correlations between change and response probabilities of participants for fine-grained and coarse-grained segmentation were computed (see Figure 6) as a function of time.

Figure 6A and Figure 6C show that participants utilized the degree of change for detecting fine-grained event boundaries. On the other hand, Figures 6B and 6D show that coarse-grained segmentation responses were uncorrelated with the degree of change. This finding entirely contradicts that of Hard et al. (2011). This might be because of the difference between these studies in stimulus types used (PLDs vs. naturalistic stimuli) and techniques employed to quantify the degree of change

(Absolute difference between points in successive timesteps vs. change detection algorithm). This point will be enlightened in the discussion.

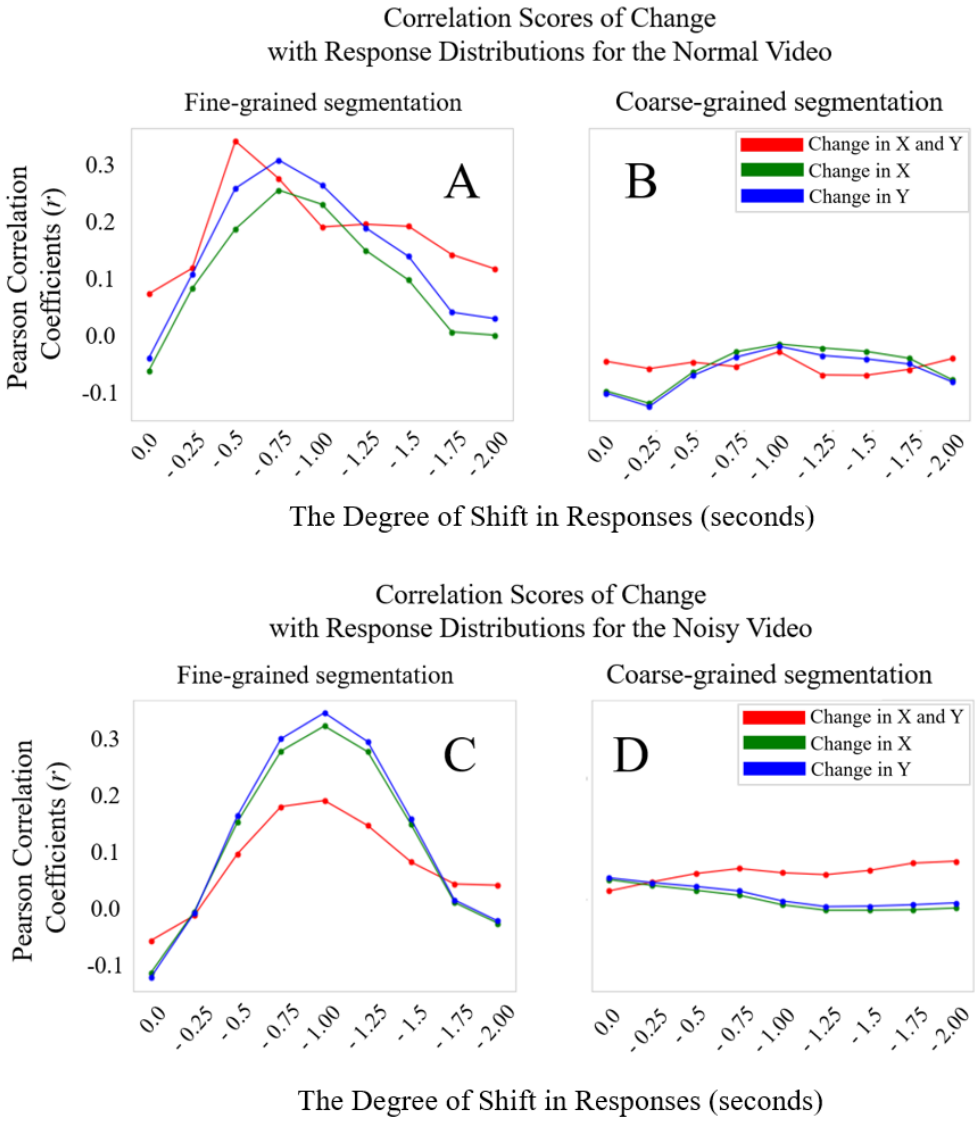


Figure 6 Event boundary responses and the degree of change

The Pearson correlation coefficients were calculated between event histograms with 0.25-second bin size and the degree of change. The unitization paradigm requires pressing a button to mark an event boundary. Certain processes such as decision making and motor modality activation for button pressing may produce a lag. For this reason, computed fine-grained and coarse-grained event histograms were shifted backward in time to reveal the genuine relationship between the absolute sensory change and event boundaries. The absolute sensory change could predict fine-grained segmentation decisions (A, C) but coarse-grained segmentation decisions were uncorrelated (B, D).

2.2.2 Computational experiments of event segmentation and learning

Now that it is stated how the ground-truth data were obtained, the training and testing procedures employed for the computational model can be explained.

2.2.2.1 Procedure

Before feeding the computational model with data, the normal and noisy datasets were subjected to a min-max normalization operation. To maintain the computational efficiency and reduce the processing time, the dataset was reduced to 4 Hz, which amounts to 1076 timesteps representing X and Y dimensions of 14 markers. The computational model was written in Jupyter Notebook by Python 3.7.6 programming language. All machine learning models were prepared by using TensorFlow-Keras (Chollet, 2015).

Twelve computational models were trained for four groups (sensory reliability x segmentation granularity). Each computational model was then tested for the same trained sequence. Hyperparameters of the model were determined for the fine-grained and coarse-grained segmentation of the normal dataset/video. The same hyperparameters were then re-used to segment noisy dataset/video to observe the effect of reduced change on the computational model's segmentation decisions. Therefore, it was aimed at testing whether the proposed model (1) captures human event segmentation decisions qualitatively and quantitatively and (2) is affected by the reduction of the rate of change.

The selected hyperparameters are given in Table B1 (Appendix B). As is described earlier, the confidence parameter θ regulates the surprise threshold φ such that higher θ makes segmented events coarser as upcoming errors are more likely to be below the surprise threshold. Another critical parameter determining the decision

of the model is the window of rolling mean, namely *w*. Determining the denominator of the rolling mean operation, *w* regulates the impact of recently received prediction error signals on φ. Higher θ and *w* values were selected for coarse-grained than fine-grained segmentation.

2.2.2.2 Results

In this section, firstly, how the proposed computational model utilizes prediction error signals for event segmentation and learning was explained. Secondly, decisions of the computational model with ground-truth data from psychological experiments were assessed.

Segmentation decisions of the computational model for a simple movement sequence were shown in Appendix C. The figure also depicts how the proposed computational model utilized prediction error signals for detecting event boundaries. The computational model computes a dynamically changing threshold value for each event model to detect their suitability for the timestep. If the current error of the event model exceeds the surprise threshold rate, the system enters the searching period to find the best event model. In each run of the computational model, the locations when an event transition occurs were accepted to be an event boundary. For both normal and noisy videos, X and Y trajectories of the given sequential information, response probabilities of participants for the fine-grained and coarse-grained event boundaries, response probabilities of trained computational models, and the accompanying absolute sensory change were given (see Appendix D and Appendix E). The mean of mean squared errors seen by event models throughout training epochs was given in Figure 7. Figures 7A and 7B represent the mean of mean squared errors received by models through training epochs. Figure 7C displays

the mean thresholds and errors received by event models in the segmentation of normal videos. In line with the EST, tolerances of event models (i.e., thresholds) determine coarse-grained and fine-grained event models.

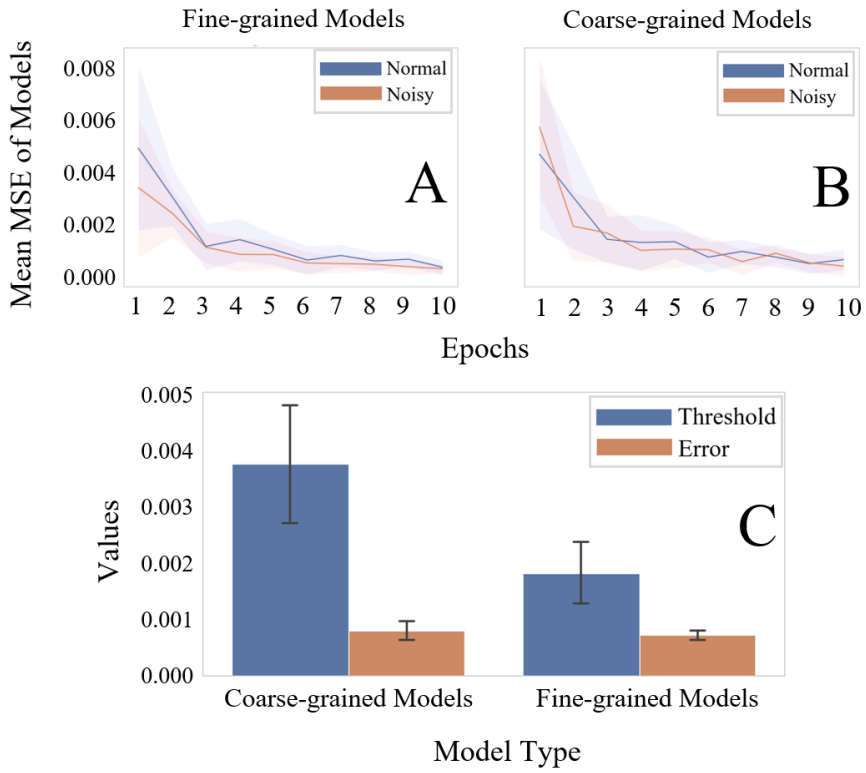


Figure 7 Training errors and thresholds of event models

(A) and (B) show training errors through epochs for fine-grained and coarse-grained event models. The Mean Squared Error (MSE) received by each event model in the online prediction phase was averaged to compute the training errors. Then, the mean errors received by each event model in each epoch were averaged. (C) represents thresholds and received errors from fine-grained and coarse-grained event models in the last epoch for the normal video segmentation. As predicted by the EST, coarse-grained event models had more tolerance to observed prediction errors. Error bars and fields represent +/- *SD*.

In line with the literature (Franklin et al., 2020), the point-biserial correlation technique was used to quantify the segmentation decisions of the computational model. Firstly, point-biserial correlation values of individual participants with respect to 1-sec event boundary histograms for the normal video were calculated, giving

participants performance distributions. Secondly, the same operation was applied for revealing the performance of the proposed computational model. Finally, to compare the proposed model with a random baseline, two control models were devised: random and change models. While the random model selects event boundaries uniformly by giving each timestep the same weight, the latter selects them by considering the absolute sensory change at that timestep.

Figure 8 presents that the proposed computational model was better than control models (sampled 12 models for each). For fine-grained segmentation, the proposed computational model received mean r = .199, where minimum r is .141 and maximum r is .27. The performance of the model was significantly more correlated with event boundary histograms than random (r = .018, z = 2.11, p = .017, one tailed) and change (r = .001, z = 2.31, p = .01, one tailed) models. The proposed model (r = .12, minimum r = .067, maximum r = .21) was also better at deciding coarse-grained boundaries than random (r = -.015, z = 1.57 p = .057, one tailed) and control (r = -.016, z = 1.59, p = .055, one tailed) models. In contrast, the correlation value of the proposed computational model with respect to fine-grained (r = .394, z = 2.48, p = .006) and coarse-grained (r = .366, z = 3.03, p = .001) ground-truth was significantly less. Note that mean values were used for statistical comparisons.

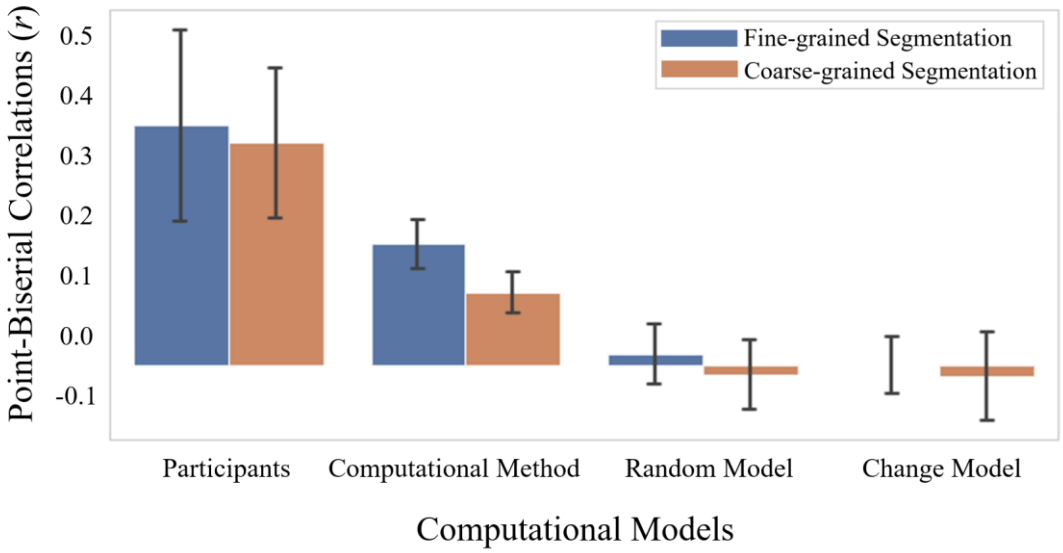


Figure 8 The correlation scores of computational models in event segmentation

The point-biserial correlation scores were computed between event segmentation decisions of participants and models with event boundary histograms produced from group data. In this way, the score of average participants in event segmentation can be compared with those of models. For example, the mean correlation of participants with the group is approximately .4 for fine-grained and coarse-grained segmentation. It is also clear that the proposed model outperformed control models. The y axis shows the mean and corresponding *SD*.

The model successfully predicts event boundaries; however, considering the temporal nature of event boundaries, a direct comparison without considering time might not thoroughly assess the model's capabilities and may hinder revealing the actual performance of the computational model. For example, participants' responses marking event boundaries might be delayed due to the time taken for decision-making or acting. In contrast, a computational model might anticipate event boundaries as it does not have such time lags. Due to this fact, for a thorough comparison, performances of computational models should be calculated by considering possible time lags, which is similar to the analysis revealing the relationship between the absolute sensory change and response probabilities (see Figure 6).

Figure 9 shows the time-dependent performances of computational models. To determine whether the proposed model anticipates event boundaries before participants' responses, decisions of computational models were shifted forward in time up to 2 seconds (8 frames).

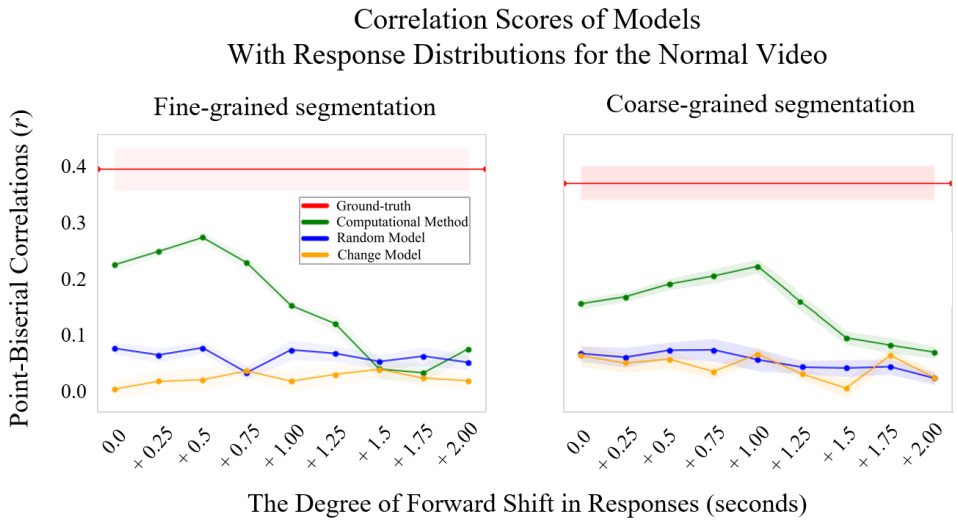


Figure 9 Time-dependent correlations of computational models

The unitization paradigm needs participants to decide which timestep is an event boundary. Specific processes such as decision making and activating motor modality for button pressing may take time. To take this fact into account, event boundaries decided by the computational model were shifted forward in time. In this way, point-biserial correlation coefficients were computed for each model. The figure shows the mean number of correlations calculated for each computational model with corresponding standard errors of the means (+/- *SEM*s). (A) and (B) show that the proposed computational model was better than control models and reached a comparable performance with humans in fine-grained and coarse-grained segmentation tasks.

The performance of the computational model in fine-grained segmentation reached its peak point at a 0.5-second shift with r = .254 (r_{max} = .32, r_{min} = .188) when the mean r of participants was .394 (see Figure 8). These two correlation values were not statistically different (z = -1.81, p = .068, two-tailed). Moreover, the best version of the model (r_{max} = .32) received a more similar performance with the

population (z = -0.98, p = .32, two-tailed). On the other hand, for coarse-grained segmentation, the performance of the computational model was maximized after a 1-second shift (r_{mean} = .196, r_{max} = .256, minimum r_{min} = .06), when the mean r of participants was r = .366, which were differed significantly (z = -2.14, p = .032, two-tailed). Nevertheless, the best coarse-grained segmentation performance (r = .256) was very comparable to those of participants (z = -1.40, p = .158, two-tailed). Overall, these results demonstrated that the proposed model reached a considerable event segmentation performance and could anticipate when an event segmentation occurs.

It can be understood from Figure 6 that the performance of participants in coarse-grained event segmentation is not explainable by change. A similar analysis can be employed to investigate the relationship between responses of the computational model and the absolute change (see Figure 10). The figure shows that the computational model is responsive to absolute changes seen in the video. Its fine-grained segmentation decisions were more explainable by absolute change than its coarse-grained segmentation decisions. In this respect, it can be said that it is roughly similar to the discrepancy shown by participants between fine-grained and coarse-grained segmentation (see Figure 6). One peculiar feature of this relationship is that absolute change in the X dimension was nearly not correlated with responses of the computational model, which might be because of the change produced in the X dimension compared to the Y dimension. Remember that the driving factor of the computational model is prediction error signals, which combine prediction errors in both X and Y dimensions. Suppose the amount change, and consequently, the prediction error is less in the X dimension. In that case, it does not accompany event segmentation boundaries because the produced error due to the change in X

dimension does not exceed the threshold, which is computed by the combination of prediction error signals received from two dimensions.

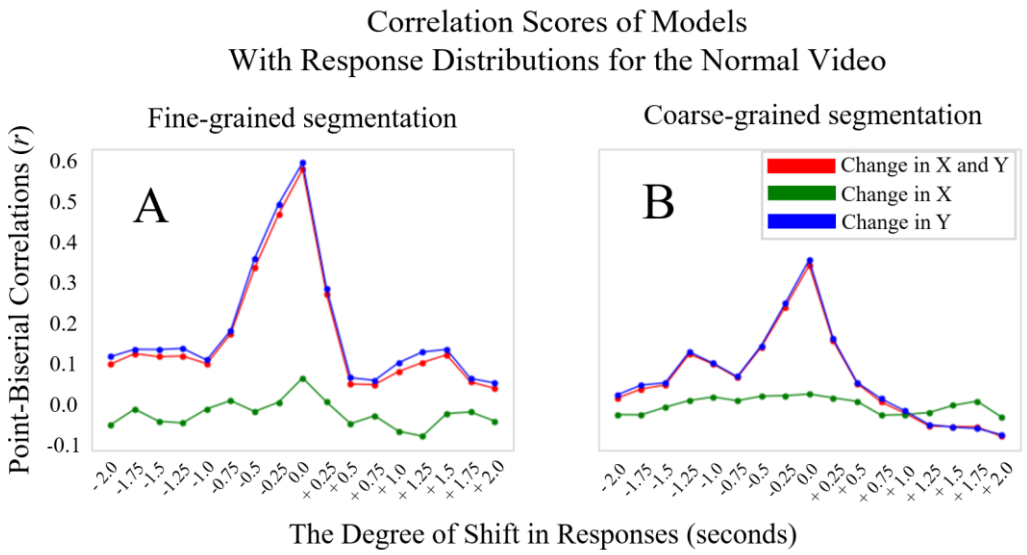


Figure 10 The relationship between responses of the proposed model and change

The figure shows time-dependent correlations between absolute sensory change and responses of the proposed model for fine-grained (A) and coarse-grained (B) segmentation. The coarse-grained segmentation decisions of the proposed model were less explainable by absolute change than its fine-grained segmentation decisions. In comparison with the decisions of participants, the higher correlation scores were received when decisions were not shifted, which shows that the computational model reacts to the change immediately, as predicted.

So far, the capabilities of the proposed computational model in the segmentation of the normal video have been discussed. This study further aimed to examine whether decisions of the proposed computational model would be biased by the quality of sensory input (video type). Note that the number of segments produced for the noisy video was less than those produced for the normal video. Figure 11 displays the number of event boundaries detected by the computational model along with the ground-truth data. Here, labels in the X dimension show conditions from which the number of responses was received. Figure 11B shows the number of responses when the computational model is trained and tested for the same video.

Figure 11B shows that the sensory information quality biased the decisions of the computational model in a reverse trend observed in human data (see Figure 11A) for fine-grained segmentation but not for coarse-grained segmentation, which is contrary to the expectations.

Note that retraining the computational model for the noisy video might not be the best way to test whether reduced change or sensory reliability bias the computational model in the same way as humans. Participants might have tried to use their already acquired event models (for example, walking), have higher tolerances for the upcoming prediction errors, and make prediction and event segmentation. In line with this idea, the computational model trained for the normal video was also tested for the noisy video to see whether its decisions would be biased by the change in sensory quality (see Figure 11C). It can be seen from the figure that when the computational model trained by the normal but tested by the noisy video displays not exact but a similar bias with humans.

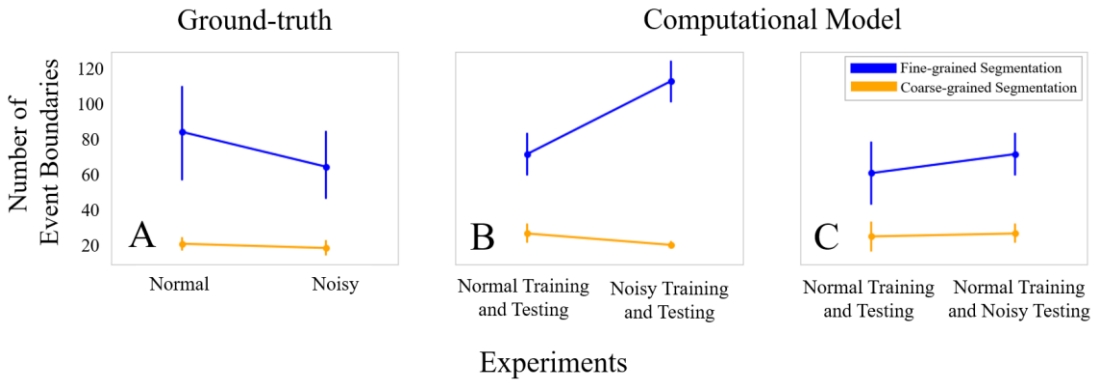


Figure 11 The responses for normal and noisy videos

The figure shows the mean number of event boundaries detected by computational models and participants with corresponding *SEMs*. Note that *SEM*s were not corrected for individual variability. (A) displays the mean responses of participants for normal and noisy videos. Fine-grained responses of participants were reduced from the normal to noisy video. (B) and (C) display different test runs. (B) represents when the computational model is retrained for the noisy video, while (C) illustrates when the computational model is tested for the noisy video.

Along with the comparison based on the number of boundaries detected in videos, event segmentation decisions of the computational model for the noisy video were evaluated by ground-truth data. Results revealed that the performance of the proposed model trained for either normal or noisy videos was better than control models (see Figure 12). Figure 12A and Figure 12C represent that computational models trained for fine-grained segmentation reached their peak points when their responses shifted 1-sec. On the other hand, Figure 12B and Figure 12D illustrate that computational models trained for coarse-grained segmentation were better than control models. However, those performances were not comparable to the results expressed so far. It should be emphasized that hyperparameters of the computational model were selected for the normal video, not for the noisy video.

Remember that while preparing videos, coarser events taken from the KIT Motion-Language Dataset (Plappert et al., 2016) were added to one another by interpolation. Computational models might exploit those interpolated trajectories. For this reason, analyses on event segmentation were also performed without them (Appendix F and Appendix G).

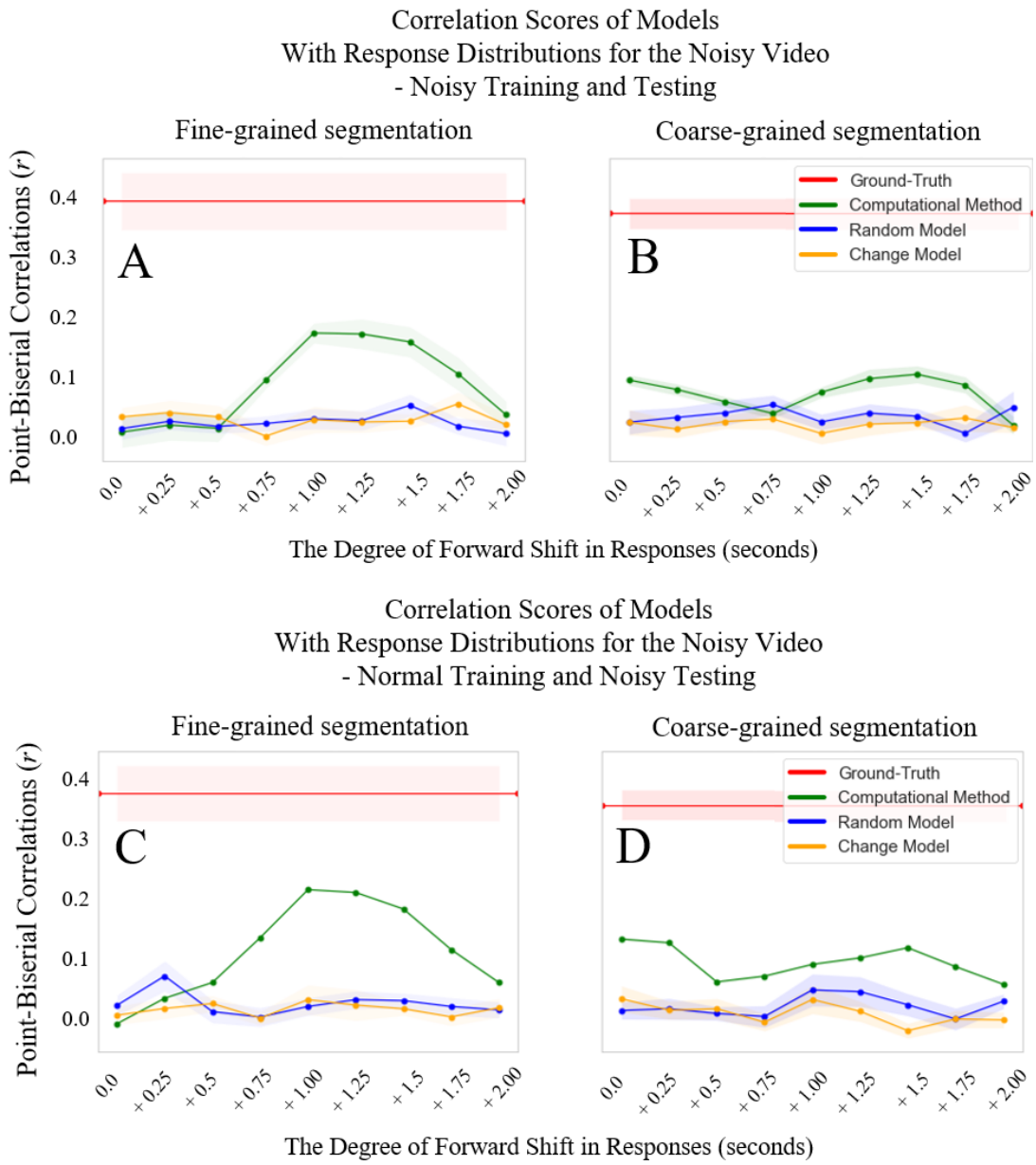


Figure 12 The performance of the model in noisy video segmentation

(A) and (B) show the performance of the model when it was trained and tested for the segmentation of the noisy video. On the other hand, (C) and (D) display the performance of the model when it was trained for normal and tested for the noisy video. Point-biserial correlation values were calculated between event boundary histograms as ground-truth and event boundary decisions of models. Event boundary decisions of models were shifted forward in time to reveal time-dependent correlations. Note that points refer to mean values, and fields represent *SEM*s.

2.2.3 Summary of event segmentation experiments

In short, to verify the computational model with ground truth received from human data, a dataset was prepared by PLDs expressing natural human behaviors such as

jumping, push-upping, and searching for an object. Event segmentation and change were correlated in the literature (Hard et al., 2011), and reduced sensory reliability was suggested to lead to coarser event segments. Another dataset (noisy video) with a reduced rate of change was created to test this idea as well as generate a validation set for the proposed computational model. Participants were recruited to segment videos generated by two mentioned datasets in the shortest (fine-grained) and the longest (coarse-grained) possible way. At the same time, hyperparameters of the computational model were determined for fine-grained and coarse-grained segmentation.

Psychological experiments demonstrated that change was a crucial parameter for the detection of event boundaries. As the rate of change was reduced, the perceived length of events extended, and the number of event boundaries perceived by human observers decreased. This conformed to the results found by Hard et al. (2011) and predictions of this study about sensory reliability.

Computational experiments confirmed that the proposed computational model (r = .199 and r = .12 for fine-grained and coarse-grained segmentation, respectively) was better correlated with the ground-truth than control models (see Figure 8). Moreover, as time shifted forward, correlation scores of the computational model (r = .254 and r = .196 for fine-grained and coarse-grained segmentation, respectively) with the ground-truth data (r = .394 and r = .366 for fine-grained and coarse-grained segmentation, respectively) increased. This might be because the computational model does not have lags due to its very nature (see Figure 9).

Overall, the study demonstrates that the proposed computational model captured the essence of event segmentation by showing similar event segmentation behaviors with people. The performance of the model was comparable to those of

participants in terms of number and locations of boundaries and affected by the noisy video in such a way that it showed not exact but comparable bias against the reduction of change.

2.3 Event representation experiments

Psychological and computational experiments of event segmentation demonstrated that the model could capture event boundary judgments of human observers. This section discovered a mostly uncovered topic, which is similarity judgments between events.

This study had two primary aims. The first aim was to examine event similarity judgments of participants. If the similarity is a valid property from the perspective of people, then the similarity relationship between extracted segments can be exploited to resolve the central problem of the proposed computational model, which is the independence of event models. To test this, one can check whether similarity judgments of people correlate with one another. The second aim was to explore whether event segments discovered by the computational model can be represented in a latent representational space in a meaningful way.

2.3.1 Representations and similarity

Representations are extremely functional theoretical constructs in studying cognition. They refer to so-called mental objects with semantic properties and specific relationships between other mental objects (Pitt, 2020). It is thought that representations have an essential place in cognition as they provide a basis for categorization and, consequently, generalization. The relationship between two representations can be approximated by the distance/similarity between them in the

representational space. The distance/similarity between two representations can generate a representational space on which representations are located (Shepard, 1980). Due to this two-way relationship, the similarity is an invaluable metric to be employed in investigating how the system organizes knowledge and which bases it uses. Therefore, artificial intelligence and cognitive psychology have been using similarity judgments of systems and people to understand how machines and people represent information.

Representation learning is one of the core research programs in artificial intelligence, learning useful and representative information from low-level sensory data (Bengio, Courville, & Vincent, 2014). For instance, a prominent machine learning model, deep neural networks, can learn distributed and semantically meaningful data representations by mimicking complex cognitive abilities (Bengio et al., 2014; Urban & Gates, 2019). They generate more and more abstract representations of data as the number of hidden layers increases (Urban & Gates, 2019). The similarity between representations of a deep learning model can be found by a similarity or distance metric (i.e., Euclidean distance or cosine similarity) and exploited to capture semantic and categorical relationships between represented entities. For example, the semantic relationship between words and sentences (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013; Rogers & McClelland, 2005), objects (Deselaers & Ferrari, 2011), scenes (Eslami et al., 2018), and episodes (Rothfuss, Ferreira, Aksoy, Zhou, & Asfour, 2018) were captured with the help of the distance/similarity between representations learned by a deep learning system.

The formation of representations is also a fundamental subject area in cognitive science. They give researchers a clue how humans organize knowledge, generalize between instances, and make analogical transfers (Blough, 2001;

Nosofsky, 1992; Shepard, 1987; Tversky, 1977). Unsurprisingly, one way of achieving human mental representations is investigating human similarity judgments that catch the semantic relationship between represented units (Shepard, 1980, 1987; Shepard & Arabie, 1979). The role of representation and similarity judgments in artificial intelligence and cognitive science suggests that the notion of representation is valuable for comparing people and machines.

Recent research adopting this approach demonstrated that representations generated by image classifiers could approximate representations of objects (Peterson, Abbott, & Griffiths, 2018). Specifically, the research showed that image classifiers could explain a high proportion of total variance in human similarity judgments between objects and predict human categorization performances. Utilizing one-layered neural networks, further research extracted core dimensions of objects by learning from human similarity judgments and demonstrated that perceptual and conceptual properties of objects might be embedded into an interpretable low-dimensional space (Hebart, Zheng, Pereira, & Baker, 2020).

Representations of entities and their similarity judgments are intertwined such that one can find out the former from the latter and vice versa. In the context of events, is it possible to regard events as spatio-temporal objects holding a similarity relationship? If it is, then event similarity judgments can be used to compare representations formed by machines and humans.

2.3.2 Event representations and similarity judgments

Event representation literature is very rich and represents a diverse set of studies from cognitive sciences. The literature encapsulates research on predictive invocation of event representations (Blom, Feuerriegel, Johnson, Bode, &

Hogendoorn, 2020), prospective coding and producing events (Schütz-Bosbach & Prinz, 2007), similarity judgments over events depicted as sentences (Day & Bartels, 2008), computational modeling of brain activation patterns of events expressed by sentences (Wang, Cherkassky, & Just, 2017), scene and event-based mental scenarios (Sheldon & El-Asmar, 2018), event representation capabilities of children (Fivush, Kuebli, & Clubb, 1992), causality and continuity in representing events (Kominsky, Baker, Keil, & Strickland, 2021). In computational modeling, recent research developed computational models using event representations for story generation (Shen, Fu, Deng, & Ino, 2020) and learning event representations in graph embeddings (Dias & Dimiccoli, 2018; Dimiccoli & Wendt, 2020). In contrast, despite the importance of event representations, the similarity between events is a mostly concealed study area associated with action categorization.

Studying similarities between actions, researchers asked participants to arrange videos depicting actions according to their perceived similarities (Kriegeskorte & Mur, 2012; Tarhan, de Freitas, Alvarez, & Konkle, 2020; Tarhan & Konkle, 2018). This way, they investigated the degree to which different features explain action similarity judgments of participants. According to the literature, high-level semantic features such as the definition of action explain similarity judgments of participants well compared to low-level features dependent on visual shape features (Tarhan et al., 2020; Tarhan & Konkle, 2018). This line of research requires the involvement of fine-grained and coarse-grained action units as these two types might depend on qualitatively different sensory and conceptual features (Hard et al., 2011, 2006; Hard & Tversky, 2003; Kurby & Zacks, 2008; Zacks et al., 2007).

The current study leans on PLDs representing events without the pictorial depiction; thus, it can be counted as an action categorization or identification

research similar to those of Tarhan et al. (2020; 2018). The dissimilarity between this study and Tarhan et al. (2020; 2018) is that the current study involves the similarity judgment of both fine- and coarse-grained action units and does not aim for searching features predicting action similarity judgments of people best. Instead, a second online psychological experiment was designed to receive similarities between events. Those event similarities were used to find out (1) whether the notion of similarity is meaningful in the context of event cognition and (2) whether the proposed model captured event similarity judgments of people.

In particular, the performance of the proposed computational model in this task would be a sign (1) for validating the model as a model of human event segmentation and show (2) whether a similarity-based metric can connect unique event models generated by the computational model to one another.

2.3.3 Computational experiments of event representation

The proposed model segments a sequence of information into unique events which are not connected. Although perceptually similar parts in a sequence are tended to be represented by the same event model, they do not have to do so. The independence of event models generated by the computational model brings limitations to generalization and hierarchical segmentation of events, which are solvable by relating event models. In other words, the proposed model is based on bottom-up processing and segmentation of sensory information without the proper guidance of familiarity and knowledge (i.e., top-down influence). A meaningful latent space representing the relationship between event models can be exploited to form a top-down influence regulating bottom-up information processing.

In this section, two representation discovery techniques were developed for relating event models and their respective videos in a latent representational space, namely hierarchy-based and prediction-based techniques. Then, event similarity judgments of the proposed computational model were computed.

2.3.3.1 Hierarchy-based technique for representation discovery (HBT)

An image classifier learning to classify objects can represent each object by its neural activations. One can find the relationship between, for example, a table and a chair by feeding their respective images into a trained multi-layer perceptron to receive their activations (representations) and computing the distance between those activations (Deselaers & Ferrari, 2011). This is because the network learns generalizable and meaningful features across inputs according to the task. Even though the proposed model utilizes multilayer-perceptrons as event models, activations produced by event models cannot be compared directly as their independence make this comparison meaningless.

Here, constructing a similarity relationship between event segments, a hierarchical multi-layer perceptron model is devised to generate a map from the sensory information and an event model. Therefore, the hierarchical multilayer-perceptron grasps the relationship between usually independent event models with the help of sensory information. Formally, the hierarchical model H does (3).

$$H(S_{t-n:t}) = M_t \tag{3}$$

This way, the hierarchical model H predicts the model M at time t by receiving sensory information S with a window of n. After training H, representation R_t for S_t is received by feeding the model $S_{t-n:t}$. In order to find the distance between two segments, each sensory input S_t is represented in two-dimensional

representational space by PCA (Principal Component Analysis) (Wold, Esbensen, & Geladi, 1987) or t-SNE (T-Distributed Stochastic Neighbor) (van der Maaten & Hinton, 2008). The representation R_a for a given segment $S_{t1:t2}$ is found by $\mu(R_{t1:t2})$. The similarity between two event segments is found by (4).

$$Sim(S_{t1:t2}, S_{t3:t4}) = 1 - Euclid(\mu(R_{t1:t2}), \mu(R_{t3:t4})) \quad (4)$$

2.3.3.2 Prediction-based technique for representation discovery (PBT)

Another way of receiving representations of independent multi-layer perceptrons is to deduce their relationships from their behaviors. Each event model in the proposed computational model learns to predict a set of points in a segment. If two segments are similar, then it can be expected that two event models trained for those segments should make similar predictions, and the degree of similarity between these predictions points out the relationship between them.

In this so-called prediction-based technique, the distance of predictions of two event models is compared by the Dynamic Time Warping (DTW) (Giorgino, 2009; Ney & Ortmanns, 1999). DTW measures the distance between two time-series data when their speed varies. For example, it can capture the distance between running movements at different speeds. Formally, suppose that $S_{t1:t2}$ is predicted by M_a and $S_{t3:t4}$ is predicted by M_b, the similarity between two segments is computed by (5).

$$\begin{aligned} Sim(S_{t1:t2}, S_{t3:t4}) \\ = 1 - DTW\big(M_a(S_{t1:t2}), M_b(S_{t1:t2})\big) \\ + DTW\big(M_b(S_{t3:t4}), M_a(S_{t3:t4})\big) \end{aligned} \quad (5)$$

2.3.3.3 Procedure

Upon training the computational model for the fine-grained and coarse-grained segmentation of the normal video, event segments were extracted with the help of event boundary judgments of a computational model. Each event segment was coded by its event model and its order on the sequence. For example, if the segment is predicted by an event model named five and the third segment the model predicts, the segment is named as 5-3. After then, HBT and PBT were applied to find similarities between detected event segments.

On the other hand, for the HBT, a neural network model was trained to categorize timesteps represented by a given window. Event categories were represented by using the one-hot encoding technique. Hyperparameters decided for the neural network were given in Table B2 (Appendix B). After the neural network model was fully trained, its activations for respective frames were received and averaged to represent event segments. Two similarity matrices and latent representations were created for HBT with t-SNE and HBT with PCA by these representations.

As a baseline to evaluate proposed representation discovery techniques, two other similarity matrices were produced. Those baselines were named DTW on trajectories and DTW on change. They were the same with the PBT, except the former compared trajectories based on their actual trajectories, and the latter assessed their similarities using their changes in time.

2.3.3.4 Initial results

The computational model generated 87 fine-grained events and 18 coarse-grained event segments for the normal video in the event segmentation phase. Fine-grained

event segments were represented by 33 unique events, whereas coarse-grained event segments were represented by eight unique events.

Two neural network models were trained for mapping sensory inputs to events. Note that event labels were extracted by the event segmentation algorithm and not taken from the ground-truth data. Accuracies and losses for neural networks mapping fine-grained and coarse-grained event segments through epochs are given in Figure 13. Accuracies reached .95 and .93 in the mapping of coarse-grained and fine-grained event segments to respective sensory information. Here, an additional validation dataset was not used since the model was developed for analysis. Pearson correlation coefficients of proposed techniques and baselines were given in Appendix H.

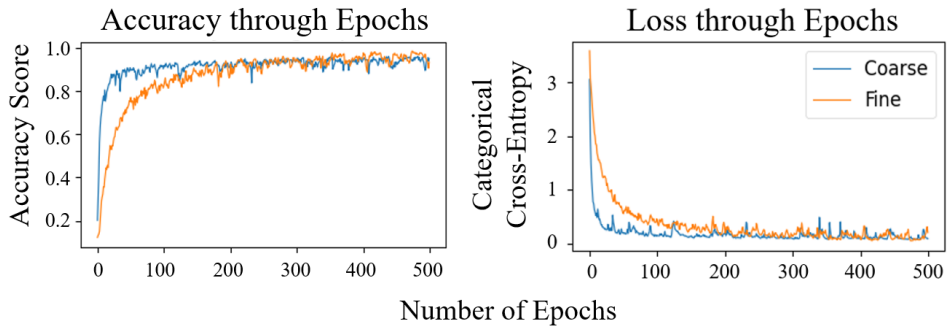


Figure 13 Accuracy scores and losses of hierarchical models

Accuracy scores and losses of the hierarchical model mapping sensory inputs to event models extracted from the computational model of event segmentation are given. It can be seen from the figures that hierarchical models successfully learned connecting sensory information to their respective events determined by the computational model.

2.3.4 Psychological experiments of event representations

An online psychological experiment including pairwise similarity judgments of events was designed to assess representation discovery techniques and discover whether similarity matrices created by HBT and PBT are meaningful. Psychological

experiments were designed by Psychopy3 (Peirce et al., 2019), and online experiments were conducted via Pavlovia.

2.3.4.1 Dataset preparation

Receiving the segmentation decisions of the computational model for the normal video, identified fine-grained and coarse-grained segments were extracted as different videos representing events. Since the current experiment was based on pairwise comparisons of videos, the number of segments was vital as it determined possible comparisons. The number of event segments extracted by the computational model in fine-grained segmentation was 87 and in coarse-grained segmentation was 18. Extreme event segments were selected for psychological experiments because using this many videos would have been unfeasible (for fine-grained 7569 and for-coarse grained 324 comparisons).

To find event segments to be used in psychological experiments, cumulative distances of each event segment with other event segments were calculated. For detecting fine-grained event segments, event segments in minimum 5% or maximum 5% were considered extreme. For coarse-grained event segments, event segments in minimum 25% or maximum 25% were considered extreme as the number of coarse-grained segments was considerably less than fine-grained segments. The exact process was applied for three techniques as they generated different distances (HBT with PCA, HBT with t-SNE, and PBT). This procedure resulted in 21 videos for fine-grained and 12 videos for coarse-grained segmentation and reduced the required comparison to 444 comparisons for fine-grained and 144 comparisons for coarse-grained event segments. Since the number of comparisons to be made is still huge, a sampling schema -explained in the procedure section- was used.

2.3.4.2 Participants

Forty-two participants (32 female, mean age 21) were recruited for the study from the Research Participation System of Boğaziçi University. Participants were mostly university students, did not have a problem with vision, and received 0.5 course credit and a chance for the 150 Turkish liras Amazon gift voucher lottery. Anonymity and confidentiality of participants were established by separating any possible information having personal references from required experimental data for analysis after the experiment was completed. Since participants were native speakers of Turkish, all experimental materials from informed consent to experimental instructions were prepared in this language. The study was consistent with the requirements of research ethics and confirmed by the Boğaziçi University Ethics Coordinating Committee (see Appendix A).

2.3.4.3 Procedure

The procedure until the entrance of the experiment was the same as the previous experiment. Each participant judged the degree of similarity between videos of fine-grained events and coarse-grained events separately. For each participant, seven fine-grained and four coarse-grained event videos were sampled. Each participant made 49 comparisons for fine-grained and 16 comparisons for coarse-grained events as different blocks, and their order was counterbalanced.

Participants were shown two videos side-by-side and had a chance to play videos by K and L for left and right, respectively. With the help of a continuous ranking slide at the bottom of the screen, they reported the degree of similarity they perceived and passed the other pair of events by pressing the space button. This

would give them a chance to check one more time to make sure of their similarity judgments (see Figure 14). Participants were asked to rate two sentences at the end of two blocks: (i) I thought that movements of points express certain human behaviors (walking, eating, or jumping, etc.), and (ii) I had a hard time comparing human movements in terms of similarities.

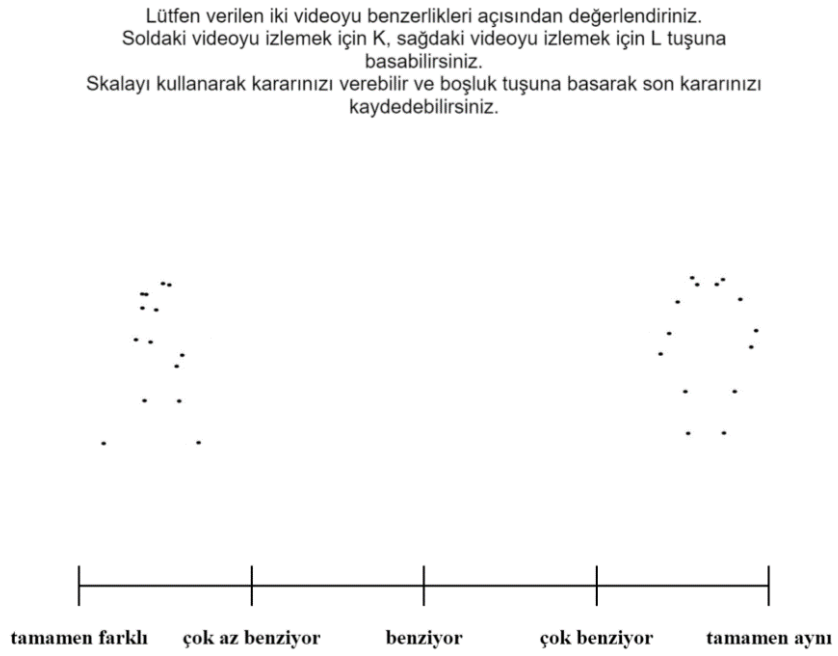


Figure 14 The screen of participants during event similarity experiments

At the above part of the figure, it is noted that “Please rate two videos according to their similarities. You can use K to watch the video at the left, L to watch the video at the right. You can use the scale for deciding and press the space button to record your decision.” The scale has the following ratings in order: completely different, a little similar, look alike, look very similar, completely identical.

2.3.4.4 Results

Upon completing the experiment, participants' subjective ratings about the task difficulty and action identification were assessed. The former could show whether similarity is a meaningful metric held between events, and the latter informed about the reliability of the experiment (Figure 15). Figure 15A illustrates that participants

sometimes had a hard time comparing actions to one another. On the other hand, Figure 15B shows that they detected and understood actions expressed by PLDs.

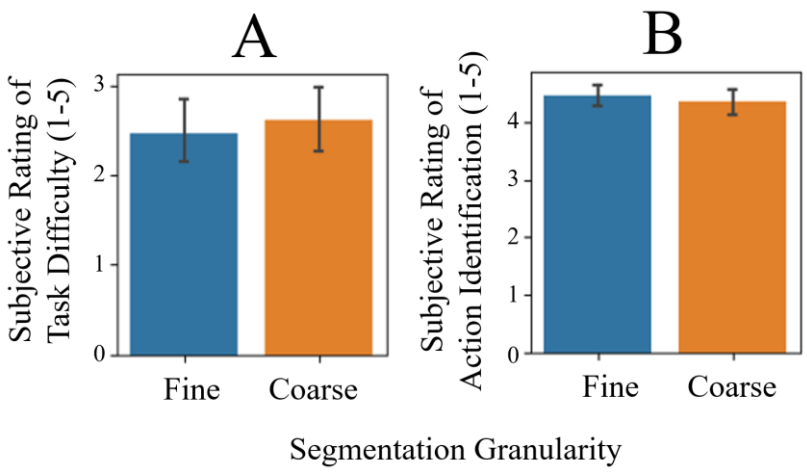


Figure 15 Subjective ratings for task difficulty and action identification success

(A) shows the degree of task difficulty. The mean is at approximately 2.5, which overlaps with "I had a hard time sometimes comparing human movements in terms of similarities." (B) demonstrates that participants understand the underlying actions represented by PLDs. Error bars represent standard deviations (+/- SD).

Participants' subjective similarity ratings were averaged out to receive the group similarity decisions. One of the most important features of similarity is that it is transitive. The order of comparison should not affect the judgments of raters. In other words, the similarity between events a and b should be the same as events b and a. Upper right and lower left side of similarity matrices were compared to one another to check this property. Two diagonal parts of fine-grained (r = .958, p = .000) and coarse-grained event similarity matrices (r = .960, p = .000) were correlated significantly. It can be concluded that transitivity as a feature was held in similarity judgments. For applying statistical tools such as multidimensional scaling and hierarchical segmentation to a given matrix, the matrix should be symmetric, meaning that two diagonal parts should include the same values. For this reason, two

diagonal parts of two matrices were averaged out, and diagonals of similarity matrices were determined to be 1.

Similar to the idea employed in the event segmentation experiments, one can receive performances of individual participants with respect to group decisions. The performance distributions of participants could be achieved by computing the Pearson correlation coefficient between each participant and the group. This type of investigation (1) revealed correlation distributions of participants, and (2) provided a chance for comparing similarity judgments of the proposed computational model with participants. Results revealed that the mean Pearson correlation coefficients of participants were .90 and .92 for fine-grained and coarse-grained event segments, respectively. Despite the report of participants in task difficulty, their correlations to the group were high.

Remember two techniques, namely HBT (Hierarchy-Based Technique) and PBT (Prediction-Based Technique), suggested for representation discovery from the proposed computational model. While the former had two results extracted by PCA and t-SNE, the latter involved comparing sensory predictions by DTW. Figure 16 displays correlations of proposed similarity techniques and baselines with group decisions for fine-grained and coarse-grained event segments. For fine-grained and coarse-grained events, similarity scores were .435 and .614 in HBT with t-SNE, .12 and .359 in HBT with PCA, .128 and .393 in DTW on predicted trajectories, .259 and .422 in DTW on real trajectories, and .149 and .405 in DTW on change.

The correlation scores of HBT with t-SNE were better than other techniques and baselines. Statistical tests revealed that correlation score of HBT with t-SNE (r = .435) was significantly higher than DTW on real trajectories (r = .259) for fine-grained units, (z = 2.973, p =.001). The same also applied for coarse-grained units

where HBT with t-SNE (r = .614) was significantly higher than DTW on real trajectories (r = .422), (z = 1.32, p = .012). Results indicated that forming internal representations from neural network activations using t-SNE yielded better results than other techniques, including baselines receiving the real data, in approximating human mental event representations. Apart from HBT with t-SNE, PBT with DTW exploiting predictions of event models reached a similar score with DTW on change in fine-grained and coarse-grained segmentation, implying that event models approximated ground-truth data properly.

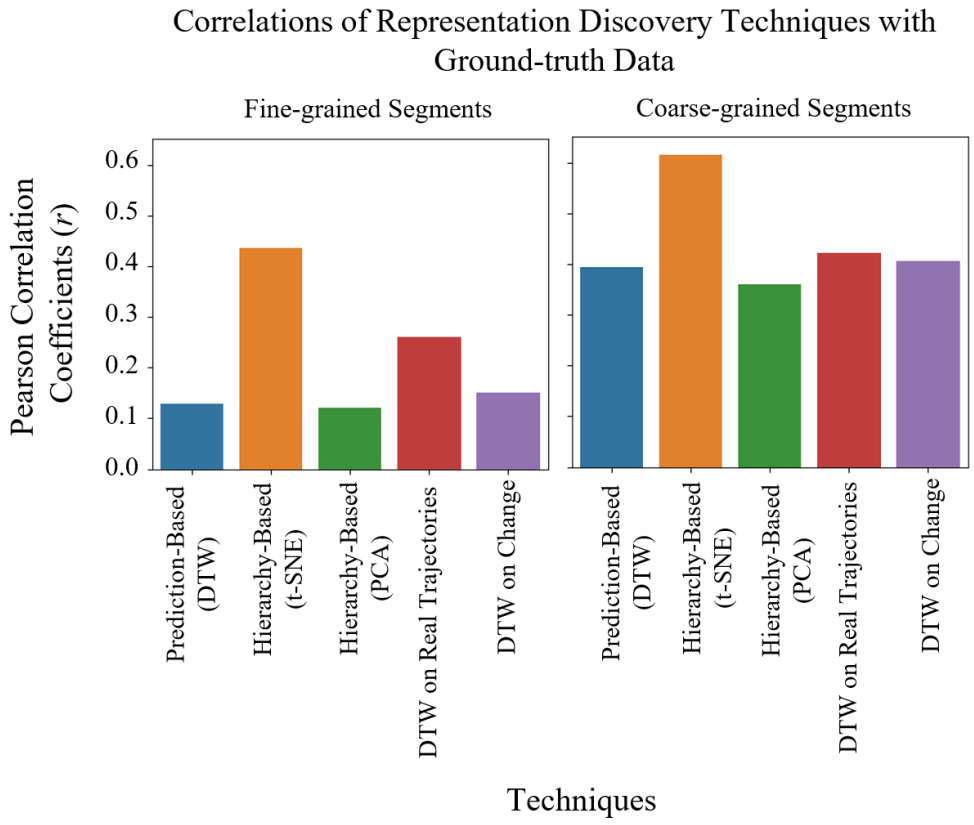


Figure 16 Correlations of representation discovery techniques and baselines

The figure shows Pearson correlation coefficients computed for assessing the relationship between similarity judgments of representation discovery techniques and humans for selected fine-grained and coarse-grained event segments. The best technique was HBT with t-SNE receiving outstanding performance compared to baseline models, namely DTW on real trajectories and DTW on change. Even though correlation scores were not approximated perfectly, they were comparable to human performance, especially for coarse-grained event segments.

So far, these results overall have shown that (1) people are capable of judging similarities between events, and (2) this property validated the proposed computational model. Similarity judgments can be converted to internal representations using self-organizing maps (SOM) (Kohonen, 1990), which is an unsupervised dimensionality reduction technique that can generate spatially organized internal representations. It can discover the semantic relationship between represented units by preserving the topological relationship between them. For this reason, self-organizing maps were used to extract meaningful representations from similarity judgments of participants and HBT with t-SNE (for selected hyperparameters and training errors, see Appendix I and Appendix J). Internal representations of HBT with t-SNE were mapped to participants' by the orthogonal Procrustes method (please look at SciPy orthogonal Procrustes method) (Schönemann, 1966). The method does not violate the Euclidian distance between represented elements.

As is mentioned, SOMs extract topological relationships between represented units by using neurons lying on a spatial map. As the distance between the two represented units increase, their similarities reduce. The resulting two-dimensional latent dimensions which represented fine-grained events were given in Figure 17. The degree of closeness is expected to represent the degree of semantic similarity. One of the first features catching attention is the closeness of representations created by the computational model and humans, aligning with the resulting Pearson correlation coefficient value (r = .435). In SOM, each neuron represents a possible cluster. It can be seen from the figure that event segments represented by two systems tend to be represented by the same or neighborhood neurons (look at 10-0 and 10-1 in Figure 17A, and 7-0 and 7-1 in Figure 17B). Moreover, SOM extracted

topological relationships between units. For instance, event segments in Figure 17A represent walking, Figure 17B walking and leaning down, Figure 17C a complete bending (i.e., crouching down), Figures 17D and 17E various kinds of hand movements. Despite this mapping between fine-grained event segment representations formed by the two systems, there are several inconsistencies. One example is representing the event segment coded as 0-1 (see Figures 17D and 17E). The computational model considered 0-1 more similar to 2-0 and 19-0, whereas humans thought that it is more like 20-0, as expected. Here the distance of the neuron representing 0-1 was very distant from neighborhood neurons, which shows the possibility that it was found quite different from the surrounding fine-grained event representations.

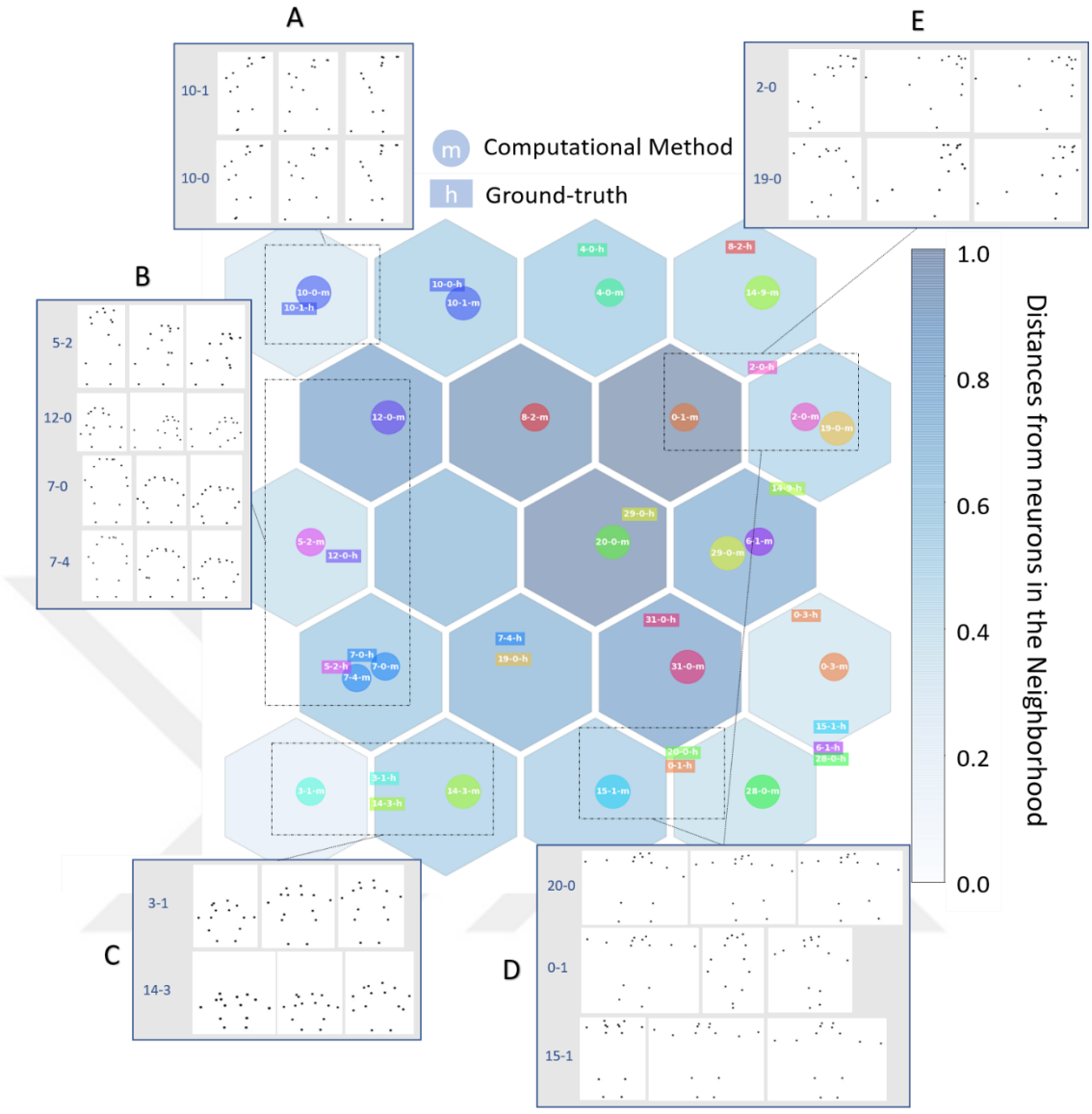


Figure 17 SOM results for fine-grained event representations

SOM was applied to similarity matrices of ground-truth and HBT with t-SNE for fine-grained event segments to extract representations in a two-dimensional map. Then, representations of HBT with t-SNE, shown by circles, were mapped to ground-truth data, shown by rectangles, by the linear orthogonal Procrustes method. The same color represents segments extracted from the same event. Event segments, in general, are represented by the same or neighborhood neurons (look at 10-0 and 10-1, 7-0 and 7-1). Moreover, SOM extracted the topology, as expected. For example, (A) represents walking, (B) walking and leaning down, (C) a complete bending (i.e., crouching down), (D), and (E) various kinds of hand movements. Despite this, there are several deviations. For instance (D and E), the computational model considered 0-1 more similar to 2-0 and 19-0, whereas humans thought it was more like 20-0.

In addition to fine-grained event segments, representations produced for coarse-grained event segments were given in Figure 18. Consistent with the degree

of correlation scores received (r = .614), locations of representations produced by two systems tended to overlap one another (see Figure 18A). In comparison to fine-grained event segments, coarse-grained event segments are rich such that they include diverse movement sequences. For this reason, it is hard to interpret topological relationships between segmented coarse-grained event units. For example, Figure 18B shows the relationships between three coarse-grained events, namely 0-3, 7-0, and 8-0. They are, in general, different event segments sharing certain similar movement parts. For example, each of them involves a bending movement. The model and humans might spot those similarities in the video. On the other hand, Figure 18C shows an interesting difference between the results of the computational model and humans. Humans found 2-0 and 2-1 similar; on the other hand, the computational model did not find the genuine relationships between 2-0 and 2-1, even if the same event model predicted them. Here, the difference might be because of the behavior of raising both hands together occurring in the event segment 2-1 but not in 2-0. This shows that humans might naturally expect that raising either one or two hands was quite similar; on the other hand, the computational model considered 2-0 more like 7-0 shown in Figure 18B.

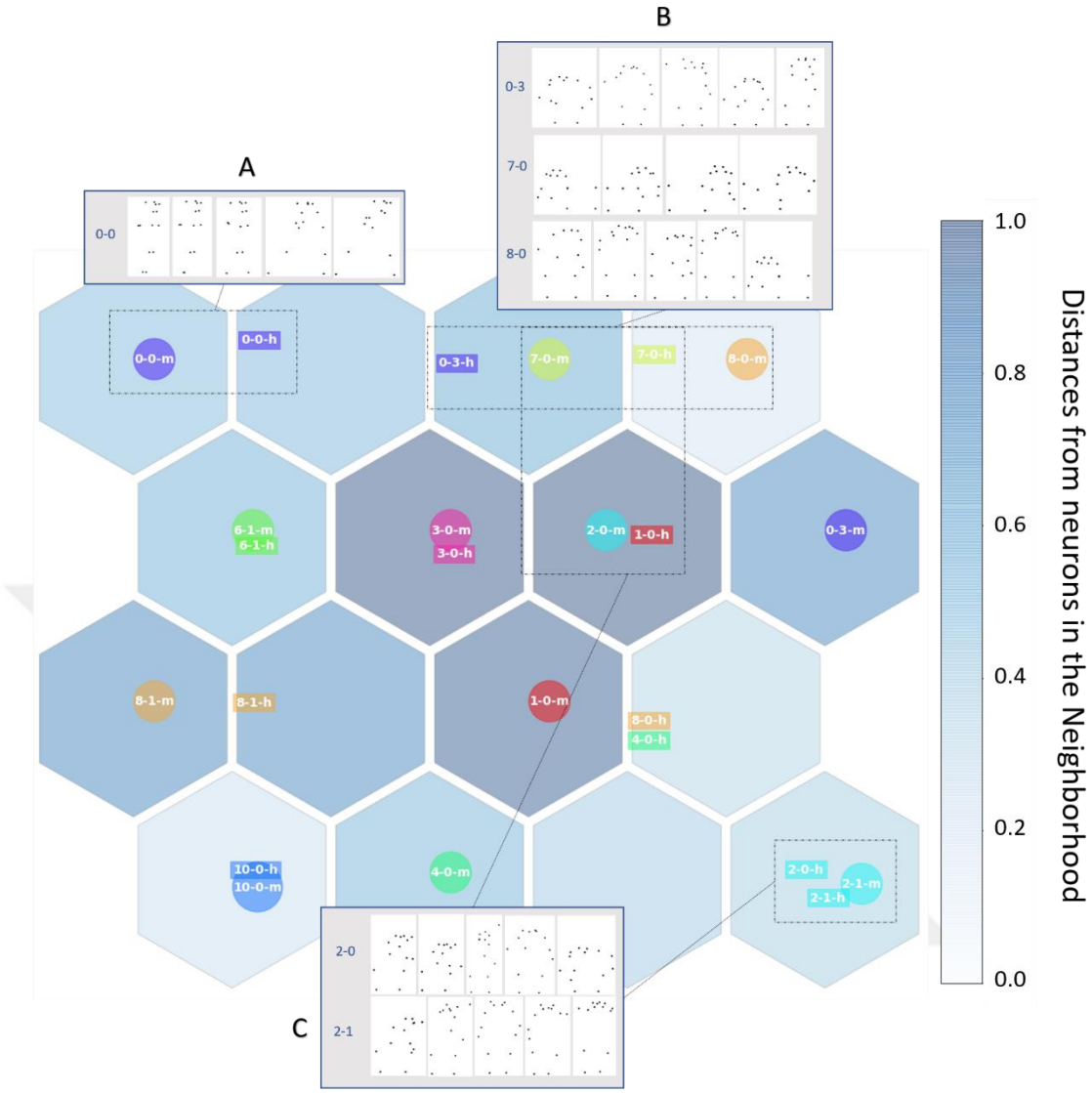


Figure 18 SOM results for coarse-grained event representations

SOM was applied to similarity matrices of ground-truth and HBT with t-SNE for coarse-grained event segments to extract representations. Then, representations of HBT with t-SNE, shown by circles, were mapped to ground-truth data, shown by rectangles, by the linear orthogonal Procrustes method. The same color represents segments extracted from the same event. It is hard to interpret topological relationships between segmented coarse-grained event units. (A) shows an instance, (B) shows the relationships between three coarse-grained events, namely 0-3, 7-0, and 8-0. They share a similar movement, which can be defined as bending. On the other hand, (C) shows a divergence between humans and the computational model. Humans, but not the computational model, found 2-0 and 2-1 similar, which might be due to the behavior of raising both hands together in 2-1 but not in 2-0. This deviation shows that humans might naturally think that raising either one or two hands should be similar.

2.4 Discussion

In this study, a computational model of event segmentation and learning was developed by considering the limitations in the literature. Some studies used stimuli having discontinuities (Metcalf & Leake, 2017; Reynolds et al., 2007), some of them was not capable of segmentation in varying granularities (Metcalf & Leake, 2017; Reynolds et al., 2007), some of them only worked in the context of interaction (Gumbsch et al., 2019, 2016, 2017), some of them did not test their model with ground-truth data (Gumbsch et al., 2019, 2016, 2017; Reynolds et al., 2007), and to best of my knowledge, the only model using human segmentation decision as ground-truth data for validation was tested one granularity level (Franklin et al., 2020).

In their model (Gumbsch et al., 2017, 2016), each event model is bound to one modality, and motor modality plays a role in event formation. These properties bring several limitations. For instance, their model cannot capture non-linear associations between different modalities and was not extended to passive observation (Newtson, 1973), which does not necessitate motor modality activation. Limitations in the literature and the model (Gumbsch et al., 2016, 2017) were overcome by developing a computational model to produce multimodal segments in varying hierarchies in a meaningful and interpretable way.

One of the first properties of the computational model drawing attention is that it received comparable correlation scores with participants. The proposed model received .199 and .12 correlation scores for fine-grained and coarse-grained segmentation, respectively. Furthermore, as time shifts forward, it received .254 and .196 mean correlation scores for fine-grained and coarse-grained segmentation. The considerable change in correlation scores as time shifts forward demonstrates that

one should be cautious in comparing artificial intelligence systems and people as they may have different characteristics making a complete comparison hard (Funke et al., 2021). In the context of event segmentation, one should consider possible time lags that can occur due to other perceptual, motor, or cognitive processes in the decisions of humans. Even though this study analyzed the performances of models by considering time lags, it was an approximation. Further research can benefit from the dwell-time paradigm that can measure locations of fine-grained and coarse-grained event boundaries without delay (Hard et al., 2011). Compared to another study that tests their model with ground truth data for one granularity level via naturalistic videos (Franklin et al., 2020), the point-biserial correlation score of the proposed model with respect to ground-truth data were high. The computational model of Franklin et al. (2020) received .143 as a point-biserial correlation. Overall, these findings demonstrate that the proposed computational model showed a considerable improvement over the literature (Franklin et al., 2020) and reached human-level event segmentation performance. However, compared to Franklin et al. (2020) using naturalistic videos, the current study used PLDs for event representation.

Correlation analyses to uncover the relationship between responses of the computational model and sensory change were conducted. When compared to absolute sensory change (see Figure 10), it can be inferred that fine-grained segmentation decisions of the computational model were driven more by absolute sensory change than their coarse-grained segmentation decisions. There was a reduction of the effect of the absolute sensory change from fine-grained to coarse-grained segmentation, as shown in the performances of human observers. In coarse-grained segmentation, the computational model may have developed a top-down

influence regulating its event segmentation decisions. However, this did not point out a complete top-down effect seen in human event segmentation decisions.

Although the proposed computational model did not receive a label on event segmentation points people make, it still needed further validation to be a promising computational model of event segmentation mimicking human perceptual and cognitive processes. One validation standard was testing the computational model for another video having less sensory change and comparing whether the computational model showed a similar bias to the video with people. In psychological experiments, the noisy video received considerably fewer event segmentation decisions. When re-trained for the noisy video, the computational model did not produce coarser segments; instead, it generated more event boundaries because the computational model adapted to the rate of change in the video and regulated its threshold values in accord with incoming errors. On the other hand, when only tested with the noisy video after trained for the normal video, the model did not increase the number of event boundaries. The latter approach is more meaningful than the former one as people might use incoming priors they developed while segmenting the noisy video. Moreover, further analyses showed that the proposed computational model received more correlation scores than control models in the noisy video segmentation even though its hyperparameters were not explicitly selected for it.

The second validation standard employed was checking whether the event segmentation model captured the similarity relationship between event segments. For this aim, an online psychological experiment was conducted to receive similarity judgments of people for events, and these judgments were compared with those of the computational model. Results revealed that people could assess similarities between events and were very reliable in their similarity judgments. Specifically, the

mean Pearson correlation score between each participant and the group was .90 and .92 for fine-grained and coarse-grained event comparisons, respectively. The proposed computational model empowered with a representation discovery technique, namely HBT with t-SNE, reached a considerable correlation score with humans (.435 and .614 correlation scores for fine and coarse event similarity judgments), although it did not achieve human performance. Moreover, received correlation scores were higher than baseline models that computed similarities between events by ground-truth data (i.e., DTW between real trajectories and change trajectories). On the other hand, PBT with DTW received a comparable correlation score with DTW on change for both fine-grained and coarse-grained events, which is sensible as event models were optimized for capturing change. It should be noted that HBT with t-SNE was dependent on the event segmentation performance of the proposed model only partially, namely in receiving labels for timesteps. In contrast, PBT with DTW was entirely dependent on predictions of event models and can be more proper for validating the event segmentation performance of the model. The success of HBT with t-SNE over proposed techniques and baselines in capturing event similarity judgments showed the possibility of a hierarchical model binding event models at a later phase of the cognitive system.

Despite this performance, HBT and this study had an importation limitation. Compared to static units such as images or words, events unfold in time. They consist of rich and diverse sequential information, which requires turning frame-based representations into event-based representations. In this study, a simple mean operation was applied to frame-based representations to extract an event-based representation. Despite its simplicity, this operation gives equal importance to each frame in representing an event, and therefore in similarity comparison. This may not

be the most proper operation as humans remember the first and the end of a sequence better than its midpoint, known as the serial position effect (Jonides et al., 2008; Murdock, 1962). Also, event boundaries were more informative for an event than within-event frames (Baldwin & Pederson, 2016; Kosie & Baldwin, 2019, 2016). Despite these limitations and suggested feature directions, the results were quite strong since the computational model did not receive ground-truth data. Instead, it autonomously captured similarity judgments of people and generated a semantically meaningful latent representation space.

The proposed model has certain shortcomings. Firstly, event models in the proposed model are distinct entities used for prediction, even though human behaviors are grouped in various clusters. That is, there might be an infinite number of versions of taking a step or jumping, and generating a model for each possibility is infeasible. In the second part of the study, different events were bound to one another in the hierarchical neural network representations. Subsequent research can examine how the relationship between event models can be fostered. Secondly, even though the proposed model can segment events in varying granularities, it could not capture hierarchical relationships between event units and use this information for coarse-grained segmentation. Although similarities between events were tried to be exploited for hierarchical segmentation of events by hierarchical clustering, it was not successful. Probably, hierarchical segmentation requires tracking higher-order statistical regularities between event units (Franklin et al., 2020; Schapiro, Rogers, Cordova, Turk-Browne, & Botvinick, 2013). Schapiro et al. (2013) demonstrated that prediction error is not necessary for event segmentation; instead, humans can learn temporal regularities between event segments and use these higher-level statistical groupings for event segmentation. It is also possible to extend the proposed model in

this way by learning the temporal regularities between event labels by a hierarchical recurrent neural network model.

Finally, despite being capable of detecting coarse-grained event boundaries, the proposed model did not develop a complete top-down effect as participants have (compare Figure 6 and Figure 10). Further research could investigate the possibility of using a dynamically formed latent representation space in a hierarchical neural network that captures relationships between events. Such a latent representation space might allow the proposed model to learn new events faster (exploiting already trained but similar models) and produce event segments based on already acquired knowledge. For the latter, principles of predictive processing can be utilized; namely, the relative reliabilities of expectation and sensory information shape what is perceived. Different confidence values of the decision in the searching period (sensory reliability) and the decision with the latent representation space (expectation) can be integrated to maintain harmony between bottom-up and top-down processes in event segmentation. Compared to data-driven models, the proposed model allows for explicit integration of systems and testing predictions of the predictive processing framework and the EST together. For example, one can extend the proposed model into a time perception model, which can aggregate prediction errors for assessing the duration of an event (Basgol, Ayhan, & Ugur, 2021; Fountas et al., 2020).

Another limitation of the proposed model is its computational inefficiency. The proposed model trains all event models to find the next event model for each searching period, which prevents modelers from using data that have a higher number of dimensions and developing more complex deep learning models for simulating event models. This is why the current study did not test the proposed

model in the segmentation of videos displaying natural activities and used PLDs to represent human behaviors. However, using PLDs brings certain limitations. For example, PLDs cannot represent human-object, human-human interaction, and changes related to background information in detail. The dimension of the RGB-coded images can be reduced by an already trained object identification model such as AlexNet (Krizhevsky, Sutskever, & Hinton, 2012) or a dimensionality reduction algorithm. In this way, the model can segment natural videos into meaningful parts. PLDs were also used as an input to the hierarchical model to maintain the correspondence. The hierarchical model used in this study was a plain feed-forward neural network discovering the similarities between fine-grained and coarse-grained event segments. In the case of using sequences of images (i.e., videos) for representing events, convolutional neural networks (CNNs) might be more advantageous than purely feed-forward neural networks. It is known that CNNs outperform feed-forward neural networks in vision-based tasks (Goodfellow, Bengio, & Courville, 2016), and they are utilized for modeling human similarity judgments in various domains (Haushofer, Livingstone, & Kanwisher, 2008; Jozwik, Kriegeskorte, Storrs, & Mur, 2017).

In the first part of the thesis, a computational model of event segmentation and learning was developed considering the literature. The computational model was validated in three ways. Firstly, its segmentation decisions for a sequence of natural human activity were evaluated. Secondly, its decisions were compared with those of humans when the sensory reliability and the sensory change were reduced. Finally, similarity judgments of the computational model were compared with those of humans. The proposed model has successfully passed all these validation tests and offers an extension to interrelated subjects.

The relationship between sensory information and expectation and their possible computational relationships were investigated with the help of the computational model. The focus was partly on the role of reduced sensory reliability in event segmentation. In the next section of this thesis, their complete relationships in determining event boundaries were examined through psychological experiments.

CHAPTER 3

PSYCHOLOGICAL EXPERIMENTS ON EVENT BOUNDARY PERCEPTION

A general overview of the content of psychological experiments was explained in the introduction. Basically, despite the existence of theoretical knowledge (De Lange et al., 2018; O'Reilly, 2013), a computational exploration (Gumbsch et al., 2017), and psychological experiments (Hard & Tversky, 2003; Hard et al., 2006; Zacks, 2004), a coherent assessment of how the system integrates sensory information and expectations in the context of event segmentation needs to be illuminated.

According to the predictive processing framework, sensory information and expectations modulate perception hand in hand with relative reliabilities. The main catchphrase to be mentioned is, when the expectation is reliable, the valence of prediction errors is reduced. On the other hand, when sensory information is unreliable, the valence of the prediction errors is reduced by giving rise to the high valence of expectations (Bastos M. et al., 2012; De Lange et al., 2018; Feldman & Friston, 2010). This relationship can be interpreted in the context of event segmentation to investigate the perception of event boundaries, as the prediction error monitoring determines their occurrence.

Hard et al. (2006; 2003) and Zacks (2004) investigated the effect of reliable expectation and unreliable sensory information. Researchers influenced participants' expectations either by showing a video multiple times (Hard & Tversky, 2003; Hard et al., 2006) or stating that movements in the videos display a purpose (Zacks, 2004). Confirming the literature (Graziano et al., 1988; Hard & Tversky, 2003; Hard et al., 2006; Markus et al., 1985; Newtson, 1973; Wilder, 1978b, 1978a; Zacks, 2004), they both found that the effect of familiarity or intention attribution makes the perceived

event segments coarser. The reduced valence of the prediction error signals resulting from unreliable expectation must have regulated perceived event segmentation in these studies. This implies that only one part of the perceptual cycle was studied (the effect of expectation reliability on sensory inputs); a coherent picture of the relationship between bottom-up and top-down processes in event segmentation necessitates examining the relative perceived reliabilities of sensory information and expectation.

From the perspective of predictive processing, reliable expectation down-grade the valence of prediction errors due to sensory input. People should observe less important prediction error signals when having reliable expectations and detect a small number of event boundaries (coarse events). In contrast, the unreliable expectation should upgrade the valence of prediction errors. When having unreliable expectations and not being confident about what they see, people should monitor more important prediction error signals and find many event boundaries (finer events). Moreover, the interaction between the reliabilities between sensory information and expectation should be meaningful. When expectations are reliable and sensory information is unreliable, people should perceive down-graded prediction error signals, leading to a reduction in the number of perceived event boundaries (coarser events). In contrast, when both expectations and sensory information are unreliable, people should perceive upgraded prediction error signals, increasing the number of perceived event boundaries (finer segments).

Since the association between bottom-up processing of sensory information and top-down influence of expectations are tightly coupled, a flexibly manipulable stimulus is required (as in similar studies, Hard et al., 2006; Hard & Tversky, 2003; Zacks, 2004). Meanwhile, the stimulus should convey meaningful sequential

information when sensory information is unreliable, noisy, and ambiguous. Here, as in the previous study, PLDs were utilized as they give outstandingly strong biological motion impressions even in the presence of noise (Bertenthal & Pinto, 1994; Giese & Poggio, 2003; Johansson, 1973; Troje & Basbaum, 2008). Along with their suitability for manipulating sensory reliability, they can be used to manipulate expectation because of the famous inversion effect, which refers to the disruptive effect of inversion on PLDs (Bertenthal & Pinto, 1994). When PLDs are inverted, people have difficulty identifying human figures (Bertenthal & Pinto, 1994; Troje & Westhoff, 2006), and the disruption in perception remains even when they know that the display is inverted (Pavlova & Sokolov, 2003). With their unique properties, PLDs allowed for investigating the relative reliability of expectation and noise on the perceived event boundaries.

In unitization-based psychological experiments, four groups of participants watched different movies with varying sensory and expectation reliabilities. The inversion effect was used to manipulate expectation reliability by showing groups either an inverted or not inverted version of PLDs (Bertenthal & Pinto, 1994; Troje & Westhoff, 2006). The noise was added to PLDs for impairing the reliability of sensory information, which inhibits participants from receiving precise and clear information.

3.1 Method

The experiment had four experimental conditions, namely sensory reliability (normal input, noisy input), expectation reliability (inverted, not-inverted displays), event granularity (fine-grained and coarse-grained segmentation), and observation order (first observation, second observation). The independent variable of the study was

the number and location of responses of participants. The experiment was prepared in Psychopy3 (Peirce et al., 2019) and conducted on an online platform named Pavlovia.

3.1.1 Participants

One hundred forty-six participants (96 female, mean age 22) were recruited for a within-subject design study. Being selected via convenient sampling voluntarily from the Research Participation System of Boğaziçi University, participants mainly were undergraduate and graduate university students. None of the participants had a problem with vision. Other details were the same as those conducted for receiving ground-truth data for event similarity judgments.

3.1.2 Experimental stimulus

As in the previous experiment, natural human behaviors were taken from the KIT Motion-Language Dataset (Plappert et al., 2016). The overall activity includes eight behaviors, which were determined to be as rich as possible. Selected behaviors were represented in PLDs format using X and Y dimensions of 16 markers (RFHD, LFHD, LBHD, RBHD, RSHO, LSHO, RELB, LELB, RWRA, LWRA, RKNE, LKNE, RTOE, LTOE, CLAV, STRN). Selected behaviors were added back-to-back through interpolating the marker positions. Here, two behaviors were added back-to-back if the distance between the end of the first and the start of the second behavior is the smallest in possible permutations. To control fine-grained segmentation decisions of participants, control behaviors that appeared in the video two times (51.16-89.33, 203.33-241.66 seconds) were included. The control behavior was the same as the one used in the previous experiment of event segmentation. Upon

preparing a base movie (6 Hz and 261-second), the movie was altered in several ways to manipulate the reliabilities of expectations participants form and sensory information videos convey.

3.1.2.1 Manipulating the reliability of sensory information

In the previous experiment, the sensory information conveyed by points in the display was manipulated two ways so that the fine-grained dynamics of behaviors were removed with the purpose of making them weak, ambiguous, and noisy (De Lange et al., 2018). In the current experiment, the same algorithm with the psychological experiment for receiving ground-truth data was used to make videos noisy; Gaussian temporal noise was added to data by window (40) and standard deviation (10). Even though this type of noise could reduce perceived boundaries in the previous experiment, it had the potential to create a confounding variable for psychological experiments as it also manipulates the sensory change (Hard et al., 2011). For this reason, the amount of sensory change shown in reliable and unreliable videos was kept the same by a simple algorithm (see APPENDIX F). The degree of change for each timestep was quantified as the absolute difference between markers in two successive frames. The algorithm changes the locations of two markers (CLAV, STRN) through time in the noisy video to match absolute changes in normal and noisy videos. While doing so, the algorithm moves markers in a virtual box created minimum and maximum values of X and Y dimensions for a particular timestep.

The resulting video had two effects on the quality of sensory information: the removal and the distortion of information. The sensory information was removed by Gaussian temporal noise such that the resulting displays lacked subtle and fine

dynamics of behaviors. On the other hand, the sensory information was distorted by markers (CLAV, STRN), undertaking matching absolute changes between normal and noisy videos. Despite this reduction in reliability of sensory information, a human figure and some behaviors were still recognizable (expectation), which was manipulated by the inversion effect.

3.1.2.2 Manipulating the reliability of expectation

The brain is a prediction machine generating expectations with the help of the current sensory data and mental models. Manipulation of expectation required not changing movement features people perceive but influencing their way of seeing the ongoing information flow. In order to manipulate the reliability of expectation, we simply turned the video upside-down by benefiting from the inversion effect, which disrupts human impression (Bertenthal & Pinto, 1994; Pavlova & Sokolov, 2003; Troje & Westhoff, 2006). In this way, the reliability of expectations that people have while observing the video was manipulated. To interfere with the human impression more, a looming operation, which shrinks and grows points in the point-light display (Hunt & Halper, 2008), was added for all videos. The algorithm that generates looming for each marker can be found in Appendix K.

3.1.2.3 Video generation

Thirty videos were generated for each between-subject group to avoid possible confounding effects, and the video to be watched and segmented by a participant was chosen randomly. In each pass of the video creation process, a video involving PLDs was generated for each group. Firstly, looming values representing the radiuses of each marker in each timestep were generated by the algorithm (see Appendix K).

Secondly, change values of normal and noisy videos -prepared by Gaussian temporal noise-, matched one another (with the help of the algorithm given in Appendix L). Finally, for inverted groups, respective normal and new noisy videos were inverted. Therefore, four different movies having the same looming history for their markers and the same CLAV and STRN positions were created. The p values and correlation scores of Pearson correlation tests between absolute change values of normal and created noisy videos were given in Figure 19. The figure shows that *p* values were between .517 and .634 and *r* values were between .942 and .957. That is, the videos shown to participants were nearly the same in terms of absolute sensory change.

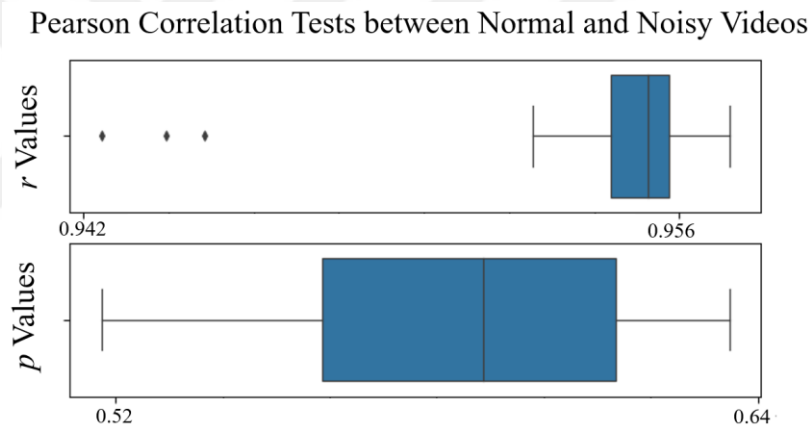


Figure 19 Pearson correlation tests between normal and noisy videos

Normal and noisy videos were created for psychological experiments to manipulate the sensory reliabilities of displays. Since event segmentation is related to change, its distortion might create a confounding variable. For this reason, an algorithm was devised for maintaining the exact absolute sensory change between videos. The figure shows the Pearson correlation test applied to reveal relationships between absolute sensory changes of normal and noisy videos. Each normal video was compared with the noisy video resulted from the algorithm. Tests established that there was no difference between normal and noisy videos in terms of absolute sensory change. Note that *p* values were between .517 and .634, whereas *r* values were between .942 and .957. That is, the two videos had the nearly same amount of absolute sensory change. Error bars represent +/- SD.

3.1.3 Procedure

The entrance of the experiment until the experimental instructions was the same as the one in the previous experiment. Each participant was randomly assigned to one of the four groups (sensory reliability and expectation reliability). The experiment started with the instruction of the segmentation granularity condition, asking participants to segment the video into either the shortest, natural, and meaningful or the longest, natural, and meaningful events by pressing the space button. The level of event granularity was counterbalanced. Participants were observed and segmented the video two times for each level of granularity. Throughout the paper, these observations were coded for maintaining simplicity, such as Fine 1, Fine 2, Coarse 1, and Coarse 2.

After the segmentation of movies in one hierarchy, participants were shown a two-choice question asking, “Which way did you segment the video you have seen?” with two possible answers “the shortest, meaningful and natural way” and “the longest, meaningful and natural way.” This question was named the attention question. After then, they were received open questions such as (i), please describe what you have seen from the videos shown during the experiment, (ii) have you ever seen similar images to those you have shown in the videos, and (iii) please mention your comments and impressions about the experiment. Following open questions, they ranked two expressions in a subjective rating scale (1-5) in accord to their perception: (i) I thought that movements of points express a human figure, and (ii) I thought that movements of points express certain human behaviors (walking, eating, or jumping). Finally, participants were requested to write their emails to be considered for the Amazon gift voucher lottery.

3.2 Results

Specific measures were adopted for maintaining the data quality. First of all, attention was paid to the participants' reports about whether the video they watched was frozen or not. Secondly, participants' answers to attention questions were examined to find out whether participants attended the experiment. Thirdly, intragroup correlations were calculated to find out diverging event segmentation decisions. Fourthly, correlations of participants to their responses (interparticipant correlations) were calculated to detect unstable segmentation decisions. Finally, outliers were detected. All reliability analyses except the freezing problem were performed based on observations because all reliability measures assessed the participants' reliability for a particular observation (Fine 1, Fine 2, Coarse 1, Coarse 2).

3.2.1 Reliability analyses

According to reports of participants, 22 participants had experienced a stuck while watching videos. Firstly, their answers were excluded from the analysis (88 observations in total). Secondly, observations that did not receive correct attention answers were detected. It turned out that 45 observations were received wrong answers, which were excluded from analyses. Thirdly, intragroup correlations were calculated to find out diverging participants. An event breakpoint histogram was computed for each group, and the group members' correlations were calculated with the group by a point-biserial correlation test. The correlation coefficients of participants to their groups were given in Appendix M. Twenty-eight observations receiving low correlation scores ($r < .1$) were excluded from the analysis. Fourthly, correlations of participants' decisions with their own decisions were calculated (like

the procedure explained in 2.2.1.5), which yielded no reliability problems. Finally, 11 observations were determined to be outliers ($z > 2.98$). These processes overall resulted in 172 observation exclusions (see Appendix N).

3.2.2 Hypothesis testing

Initially, to investigate whether the manipulation expectation worked, participants' rankings on the human impression and human figure ratings were investigated (see Figure 20). Two-way ANOVA analysis revealed that there were main effects of expectation ($F(1, 414) = 34.16$, $p = .000$, $\eta^2 = .076$) and of sensory reliability ($F(1, 414) = 35.22$, $p = .000$, $\eta^2 = .078$) on the degree that participants have seen a human figure. Specifically, Tukey post hoc tests revealed that the mean of the not-inverted group was significantly greater than that of the inverted (difference = 0.747, $t = 5.59$, $p = .001$, $r = .264$). On the other hand, the mean of noisy group was significantly less than the normal group (difference = -0.757, $t = -5.68$, $p = .001$, $r = -.2685$). That is, both observing noisy and inverted displays hinder participants from seeing a human figure. Similarly, another two-way ANOVA analysis revealed that there were main effects of expectation ($F(1, 414) = 84.51$, $p = .000$, $\eta^2 = .169$) and of sensory reliability ($F(1, 414) = 25.18$, $p = .000$, $\eta^2 = .057$) on the degree that participants have seen human behaviors. Particularly, Tukey post hoc tests further showed that the mean of the not-inverted group was significantly greater than that of the inverted (difference = 1.085, $t = 8.92$, $p = .001$, $r = .4$). On the other hand, the mean of noisy group was significantly less than the normal group (difference = -0.587, $t = -4.54$, p = .001, $r = -.217$). This means that both observing noisy and inverted displays hinder participants from seeing particular human behaviors such as walking, jumping, and eating. The primary aim of the study was to manipulate human impressions by the

inversion of PLDs. However, it turned out that noisy video disturbed participants' expectations, which should be paid attention to throughout analyses.

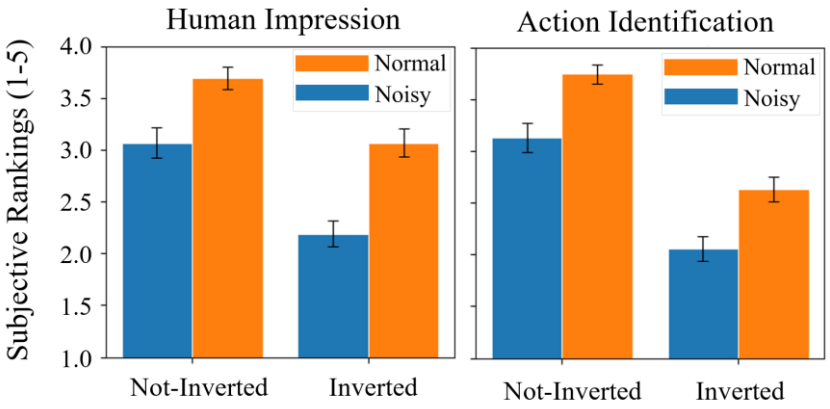


Figure 20 Participants' reports on perceived human figures and behaviors

After the experiment, participants were asked to rate two questions to understand whether manipulations worked. While the left figure shows ratings of participants for "I thought that movements of points express a human figure," the figure at the right displays their ratings for "I thought that movements of points express certain human behaviors (walking, eating, or jumping, etc.)." Error bars represent +/- *SEM*.

Here, the effects of variables on the number of responses were analyzed. Four-way between-subject ANOVA (due to excluded data) on the effects of reliability of expectation, reliability of sensory information, segmentation granularity, and segmentation order on the number of responses was analyzed. Results revealed only the main effect of segmentation granularity ($F(1, 414) = 97.89$, $p = .000$, $\eta^2 = .19$) (see Figure 21). Tukey post hoc test showed that the mean number of responses in coarse-grained segmentation was significantly lower than that of fine-grained segmentation (difference = -25.17, $t = -10.00$, $p = .001$, $r = -.44$).

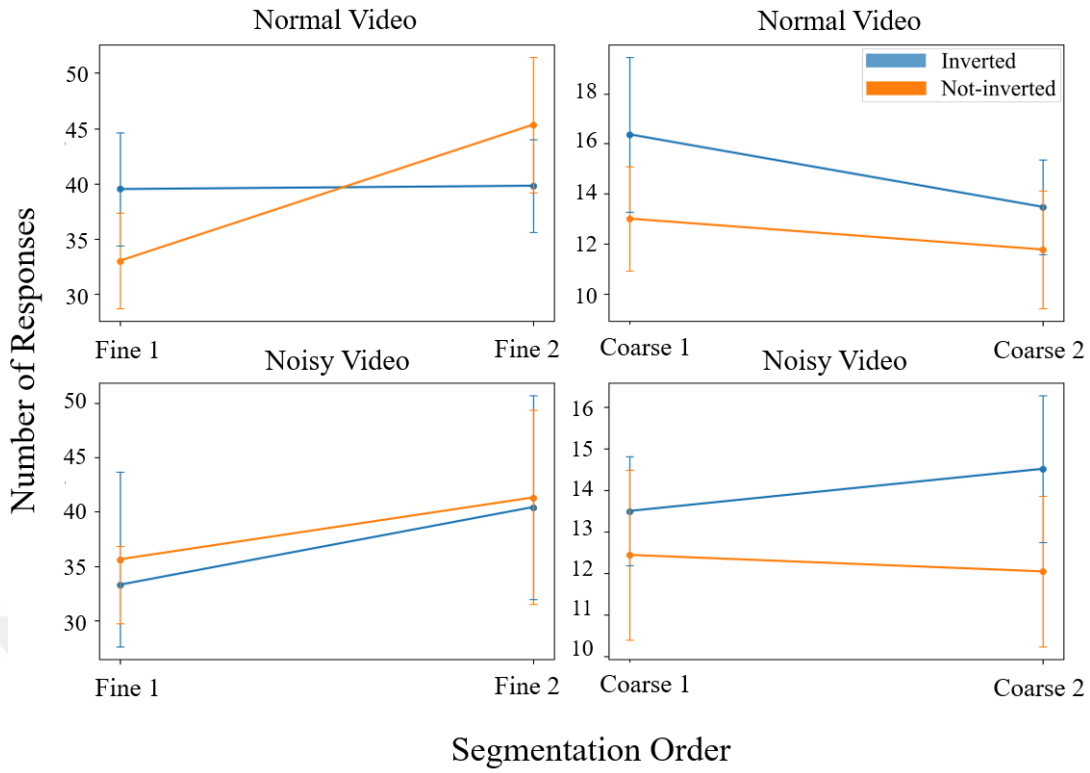


Figure 21 Number of responses in psychological experiments

The figure shows the responses of participants in different groups. The two figures at the top show responses for the normal video, whereas the two figures at the bottom display responses for the noisy video. Fine-grained segmentation responses are given at the left, whereas coarse-grained segmentation responses are given at the right side of the figure. Blue and orange colors are used for displaying the number of responses in inverted and not-inverted displays, respectively. Responses of participants revealed that there were no significant differences between groups apart from the segmentation granularity. Error bars indicate the standard errors of the mean (+/- *SEM*).

Analyses showed no significant effect of the reliability of sensory information and expectation on the number of responses. High *SEM*s seemed to hinder discovering a significant relationship between variables. The main effects of sensory reliability and expectation on the number of event boundaries were shown in Figure 22.

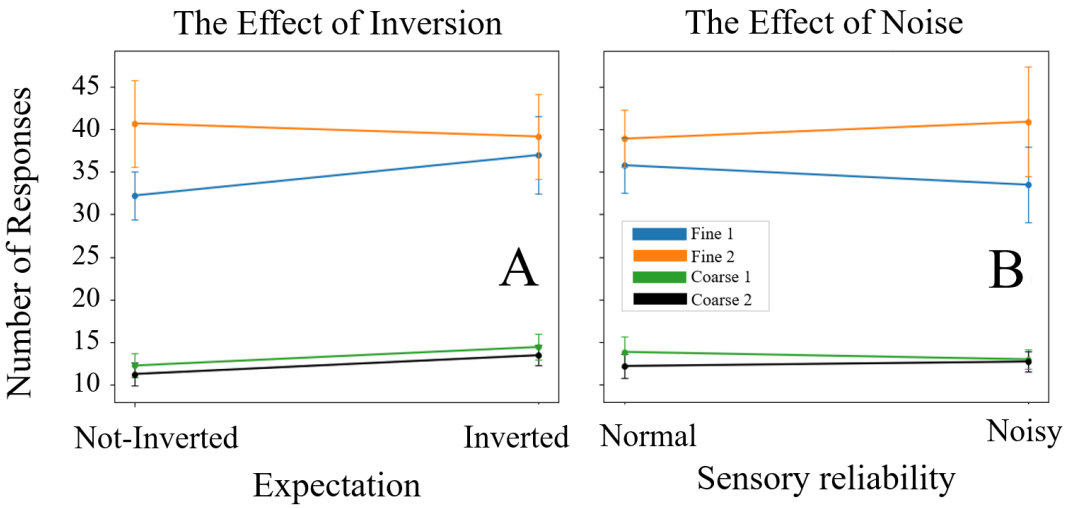


Figure 22 Effects of experimental variables on the number of responses

The figure illustrates the effects of experimental conditions on the number of responses produced by participants. (A) shows the effect of display inversion, whereas (B) displays sensory reliability on the responses. Trends were not as expected except that of coarse-grained segmentation in the display inversion. In line with the literature, participants tended to produce more segments when displays were inverted (i.e., dissimilar to already known). Nevertheless, one should be cautious since interactions were not statistically significant. Error bars indicate the standard errors of the mean (+/- *SEM*).

Figure 22B demonstrates that there was little or no effect of sensory reliability on the number of responses; on the other hand, Figure 22A shows that there might be an insignificant but considerable effect of segmentation order in fine-grained segmentation of not-inverted displays. The t-test revealed that participants did not perceive significantly greater fine-grained event boundaries in not-inverted displays in the first observation (M = 33.13, SEM = 2.80) than in the second observation (M = 42.83, SEM = 5.39), $t(100) = -1.59$, $p = .115$, $d = .32$. Nonetheless, the effect size, Cohen's d, was between small and medium (Cohen, 2013). Another t-test was applied to find out the effect of display inversion on coarse-grained segmentation. When displays not-inverted (M = 12.28, SEM = 1.03), although participants did not perceive significantly less coarse-grained event boundaries than

when displays were inverted (M = 14.39, SEM = 1.02), $t(198) = -1.44$, $p = .149$, $d = .20$, the test had a small effect size (Cohen, 2013).

3.3 Discussion

In psychological experiments, extracted hypotheses from the literature were not confirmed. The only statistically significant effect was due to segmentation granularity, which refers to segmenting videos into either fine-grained or coarse-grained events. Neither the main effect of reliability of expectation nor sensory reliability could be found. Also, their interactions were not significant. This was partly because of high *SEM*s due to the high standard deviations observed in the responses of groups. Subsequent tests showed a small-to-moderate effect size in two-times fine-grained segmentation of not-inverted displays and the inversion of coarse-grained segmentation. These effects were in line with the literature, claiming that top-down influence of prior knowledge and expectation made perceived event segments coarser (Graziano et al., 1988; Hard & Tversky, 2003; Hard et al., 2006; Markus et al., 1985; Newtson, 1973; Wilder, 1978b, 1978a; Zacks, 2004). However, the results were inconclusive and did not provide evidence for the hypotheses formed in the study.

One of the first limitations of the current study was high *SEM*s due to high standard deviations seen in groups' responses. There might be several reasons for high standard deviations. One reason might be task difficulty which participants complained about. For example, several participants commented that they thought they did not understand the task well. Some thought that the task was difficult as they were expected to segment a video having a meaning (i.e., inverted displays). Besides these problems, although event segmentation boundaries detected by participants

share certain similarities and are correlated to one another, it was observed that the number of perceived event boundaries is prone to individual variability. High individual variability might conceal the effects of manipulated variables, namely expectation and sensory reliability. Subsequent research could devise an experiment as a within-subject design to overcome this problem. Another source of individual variability might be the ambiguous definition of fine-grained and coarse-grained events. The current study did not train participants to stay away from shaping their decisions but devise more than one observation for each segmentation condition as a practice phase. Future research can try to find a way to clarify the meaning of fine-grained and coarse-grained segmentation for participants without shaping their decisions. Finally, another reason might be, despite methods employed, poor data reliability, preventing the study from finding an accurate picture of the population. Subsequent research can devise a similar experiment in the laboratory environment by paying attention to the problems mentioned.

CHAPTER 4

GENERAL DISCUSSION AND CONCLUSION

In this study, two sides of the perceptual cycle determining locations of event boundaries were examined. In the first part, a predictive processing-based self-supervised computational model of event segmentation and learning was developed, and the computational model was validated in three experimental settings with a significant improvement over the literature. Experiments showed that the reduced change and sensory reliability make segmented events coarser. A not one-to-one but comparable bias was observed in decisions of the computational model. Further, it was shown that the proposed computational model could extract relationships between events in a meaningful way. In the second part, two sides of the perceptual cycle were manipulated according to the principles of predictive processing, assuming the role of reliabilities (confidences) of sensory information and expectation in perceptual formation. Despite the efforts, evidence could not be found for the hypotheses put forward, except small-to-medium effect sizes pointing out the effect of expectation reliability on event boundary detection. Further research should search for other opportunities of manipulating two sides of the perceptual cycle by benefiting from the experience gained in this study.

Segmentation is an efficient strategy of the perceptual system that minimizes the dimensionality of continuous information flow by turning them into a set of discrete spatio-temporal objects. Although this strategy seems straightforward, it is regulated by complex interrelated perceptual and cognitive mechanisms and requires the combination of incoming sensory information and existing knowledge. Exploring the mechanism behind this strategy and its perceptual and cognitive underpinnings

may illuminate how organisms organize and use the information for adapting to complex environments.

APPENDIX A

ETHICS COMMITTEE APPROVAL

T.C.
BOĞAZİÇİ ÜNİVERSİTESİ
SOSYAL VE BEŞERİ BİLİMLER YÜKSEK LİSANS VE DOKTORA TEZLERİ ETİK İNCELEME KOMİSYONU
TOPLANTI TUTANAĞI

Toplantı Sayısı	:	12
Toplantı Tarihi	:	21.01.2021
Toplantı Saati	:	13:00
Toplantı Yeri	:	Zoom Sanal Toplantı
Bulunanlar	:	Prof. Ebru Kaya, Prof. Dr. Fatma Nevra Seggie, Dr. Öğr. Üyesi Yasemin Sohtorik İlkmen
Bulunmayanlar	:	Prof. Dr. Özlem Hesapçı Karaca

Hamit Başgöl
Bilişsel Bilim

Sayın Araştırmacı,
"Olay Öğrenme ve Ayırmanın Hesaplamalı Bir Modeli: Olay Boyutluluğu, Duyusal Güvenilirlik ve Beklenti Etkisi" başlıklı projeniz ile ilgili olarak yaptığınız SBB-EAK 2020/58 sayılı başvuru komisyonumuz tarafından 21 Ocak 2021 tarihli toplantıda incelenmiş ve uygun bulunmuştur.

Bu karar tüm üyelerin toplantıya çevrimiçi olarak katılımı ve oybirliği ile alınmıştır. COVID-19 önlemleri kapsamında kurul üyelerinden ıslak imza alınamadığı için bu onam mektubu üye ve raportör olarak Yasemin Sohtorik İlkmen tarafından bütün üyeler adına e-imzalanmıştır.

Saygılarımızla, bilgilerinizi rica ederiz.

Dr. Öğr. Üyesi Yasemin
SOHTORİK İLKMEN
ÜYE

e-imzalıdır
Dr. Öğr. ÜyesiYasemin Sohtorik
İlkmen
Öğretim Üyesi
Raportör

SOBETİK 12 21.01.2021

APPENDIX B

HYPERPARAMETERS OF COMPUTATIONAL MODELS

Table B1. Hyperparameters of the Computational Model

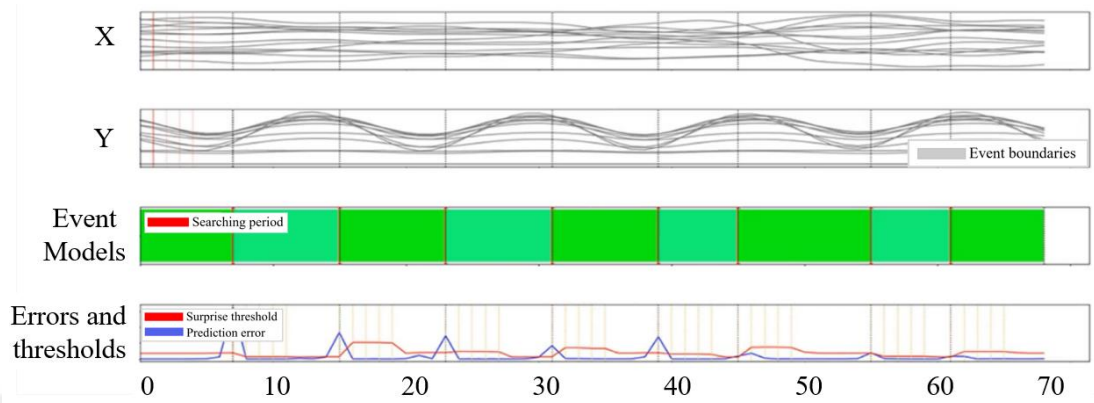
Parameters	Fine-grained	Coarse-grained
Event threshold (φ)	1.25	2.5
Error window (w)	10	30
Number of timesteps (n)	5	15
Rehearsal	100	100
Replay	2000	2000
Number of epochs	10	10
Memory range	1	1
Activations functions	Relu	Relu
Optimizer	Adam	Adam
Learning rate	0.0001	0.0001
Batch size	12	12
Hidden layers	(256, 128, 64, 64)	(512, 256, 128, 128)

Table B2. Hyperparameters of the Neural Network Trained for HBT

Parameters	
Optimizer	Adam
Learning rate	0.005
Number of epochs	500
Activations functions	Relu
Batch size	64
Hidden layers	(2048, 1024, 512, 256)

APPENDIX C

A SIMPLE SEQUENCE SEGMENTATION

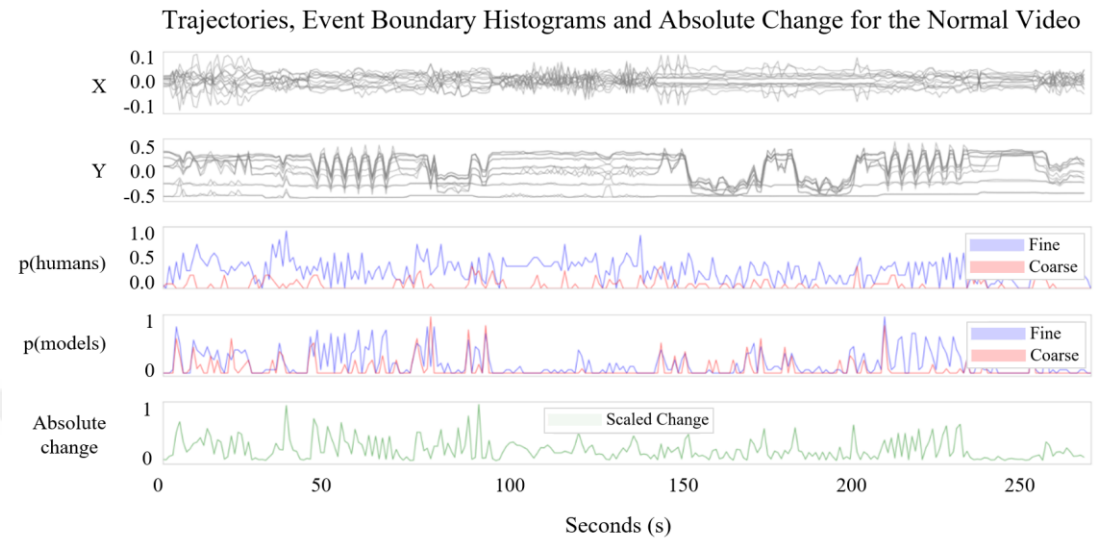


The utilization of predictive error signals of the computational model:

The segmentation results were produced by the model. The segmented behavior was defined in the dataset as "A person repeatedly picks something up from the floor and holds the object high above." The first and second figures show X and Y coordinates of dots in PLDs, respectively. The third figure shows timesteps when the model enters the search period by red lines and events by colors. The color change in the third figure shows the event boundary locations at which an event transition occurs. Finally, the fourth figure displays the surprise threshold and prediction error by red and blue colors, respectively, for each timestep. In short, the model enters the search period when the error received by the current event model exceeds its threshold value and replaces the current event model with a suitable one if necessary.

APPENDIX D

EVENT BOUNDARY HISTOGRAMS FOR THE NORMAL VIDEO

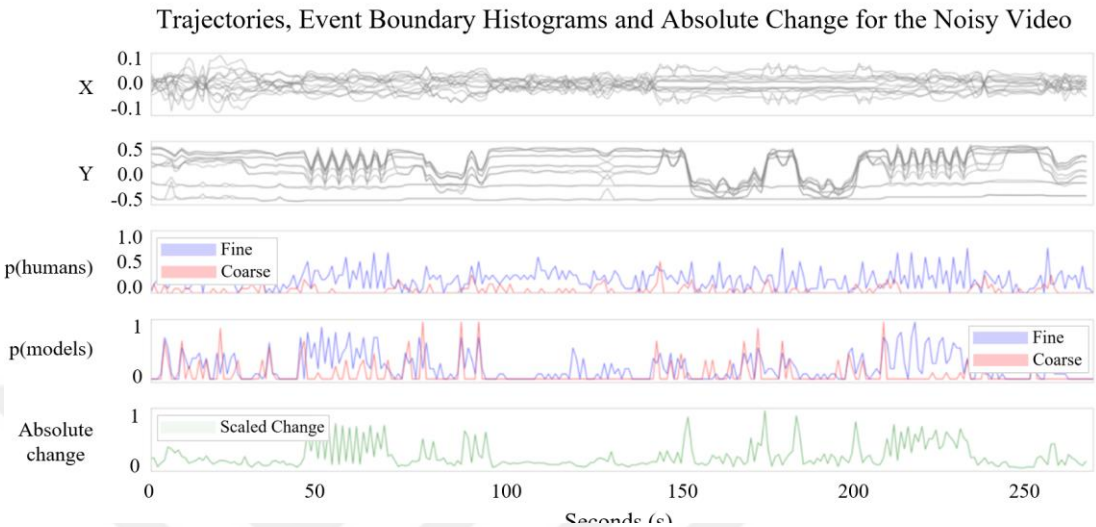


Event boundary histograms produced for the normal video:

(A) and (B) shows the X and Y trajectories of the normal video. (C) displays event boundary histograms with 1-second bin size computed to reveal participants' response probabilities as a function of time for fine-grained and coarse-grained segmentation levels and (D) does the same for computational models. (E) represents the absolute sensory change as a function of time in the normal video.

APPENDIX E

EVENT BOUNDARY HISTOGRAMS FOR THE NOISY VIDEO

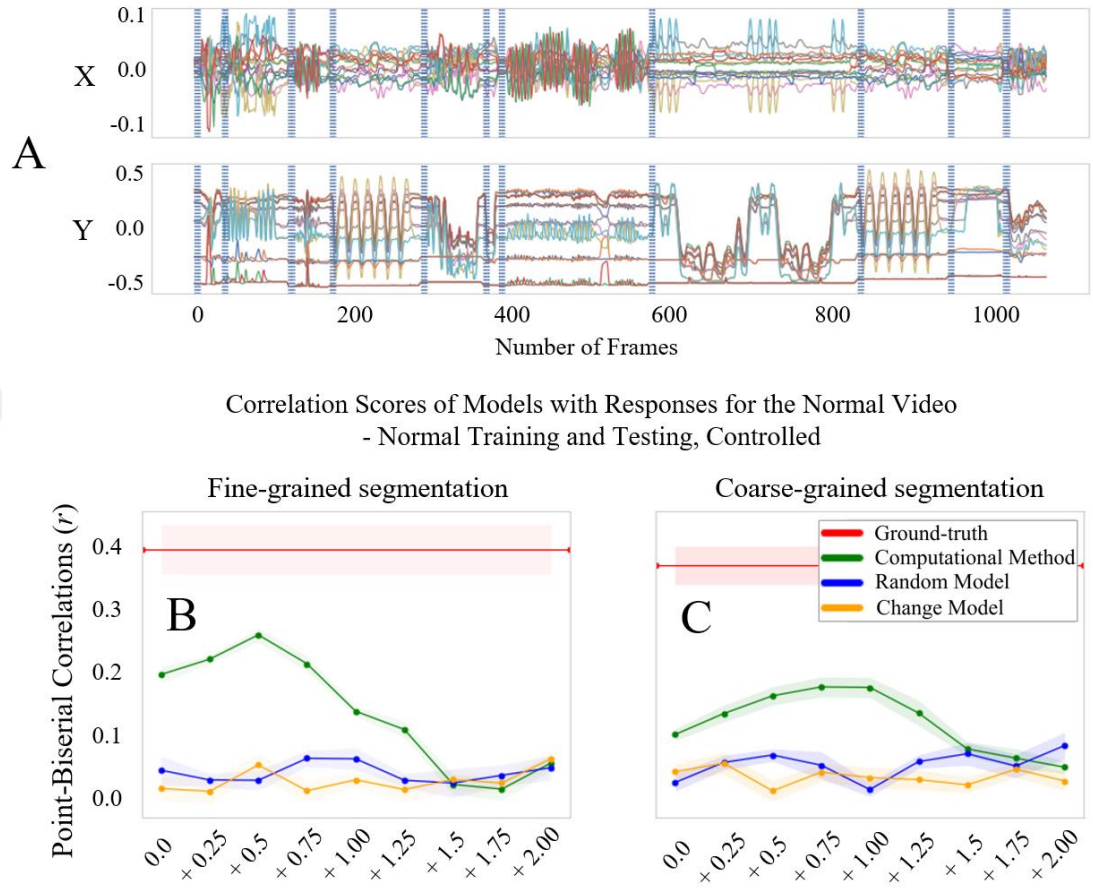


Event boundary histograms produced for the noisy video:

(A) and (B) shows the X and Y trajectories of the noisy video. (C) displays event boundary histograms with 1-second bin size computed to reveal participants' response probabilities as a function of time for fine-grained and coarse-grained segmentation levels and (D) does the same for computational models. (E) represents the absolute sensory change as a function of time in the noisy video.

APPENDIX F

CONTROL ANALYSES FOR THE NORMAL VIDEO SEGMENTATION



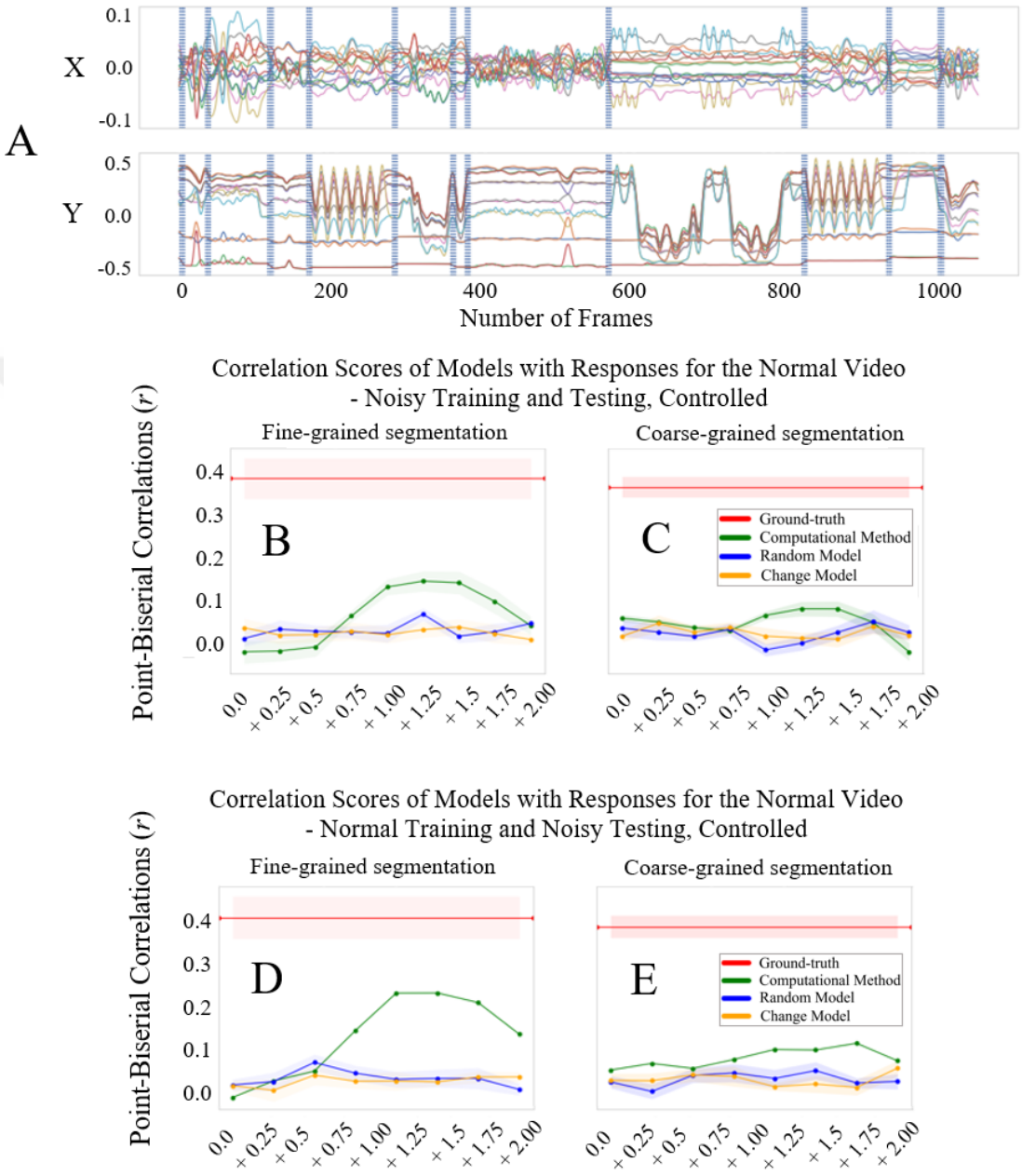
Time-dependent performances of the computational model for the normal video when interpolation points were removed:

(A) shows interpolated trajectories where coarser behavioral units were added to one other. Responses on these points were excluded from analyses. (B) and (C) show fine-grained and coarse-grained segmentation correlations. Point-biserial correlations were calculated between event boundary histograms as ground-truth and models' event boundary decisions. Event boundary decisions of models were shifted forward in time to reveal time-dependent correlations. Statistical tests revealed that performances of the model were not differed significantly for fine-grained (r_{normal} =

.199, $r_{control}$ = .171, z = 0.335, p = .737, two-tailed) and coarse-grained segmentation (r_{normal} = .12, $r_{control}$ = .06, z = 0.699, p = .484, two-tailed), when there was no time shift.

APPENDIX G

CONTROL ANALYSES FOR THE NOISY VIDEO SEGMENTATION



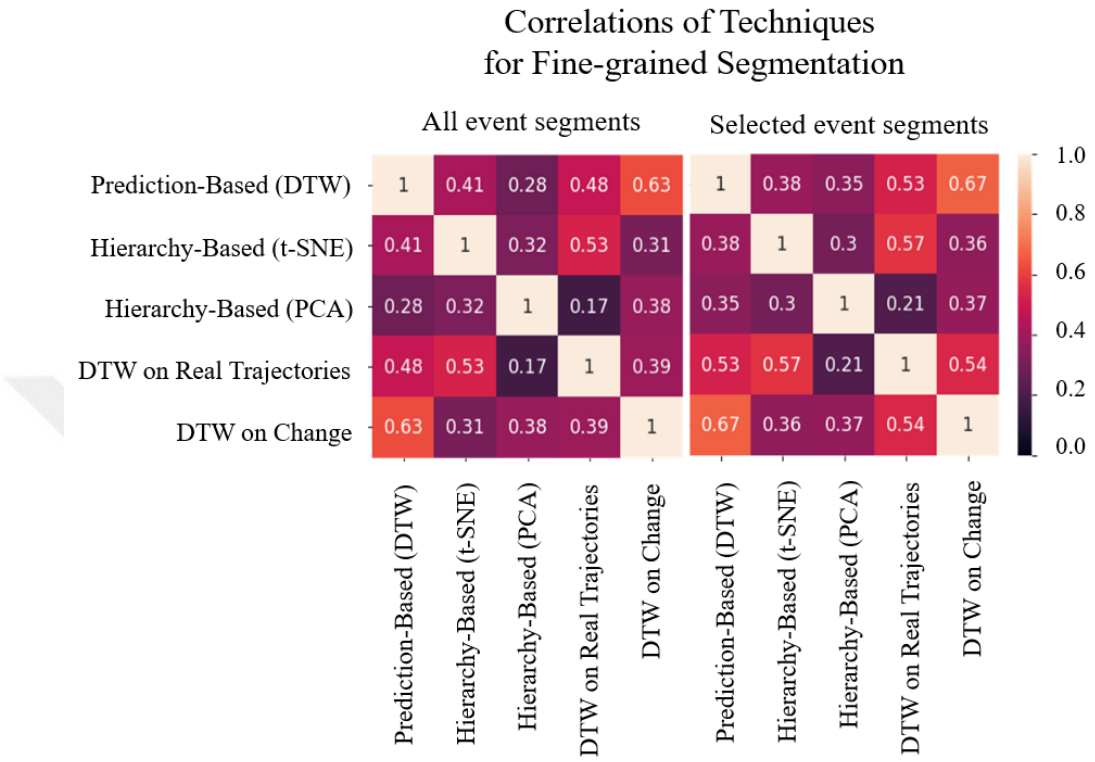
Time-dependent performances of the computational model for the noisy video when interpolation points were removed:

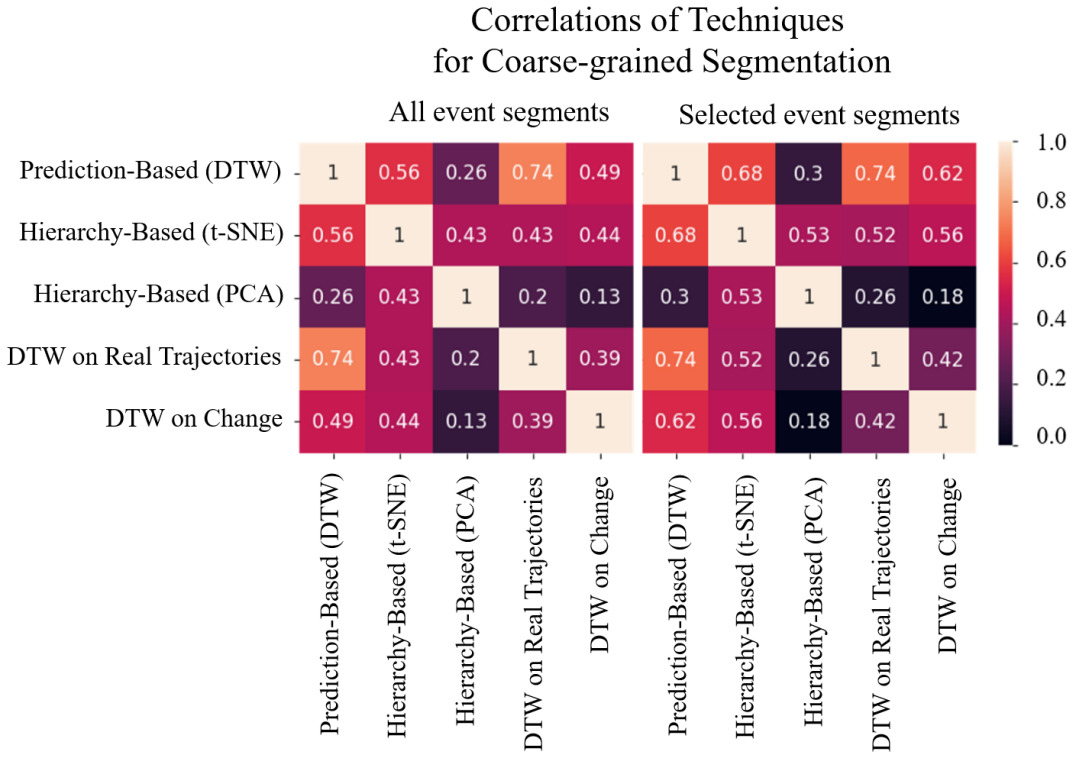
(A) shows interpolated trajectories and (B, C, D, E) show control analyses. Point-biserial correlations were calculated between boundary histograms as ground-truth

and models' boundary decisions shifted forward in time. Note that points refer to mean values, and fields represent *SEM*s. Note that hyperparameters of the model were not selected for the noisy video.

APPENDIX H

CORRELATIONS OF REPRESENTATION DISCOVERY TECHNIQUES





APPENDIX I

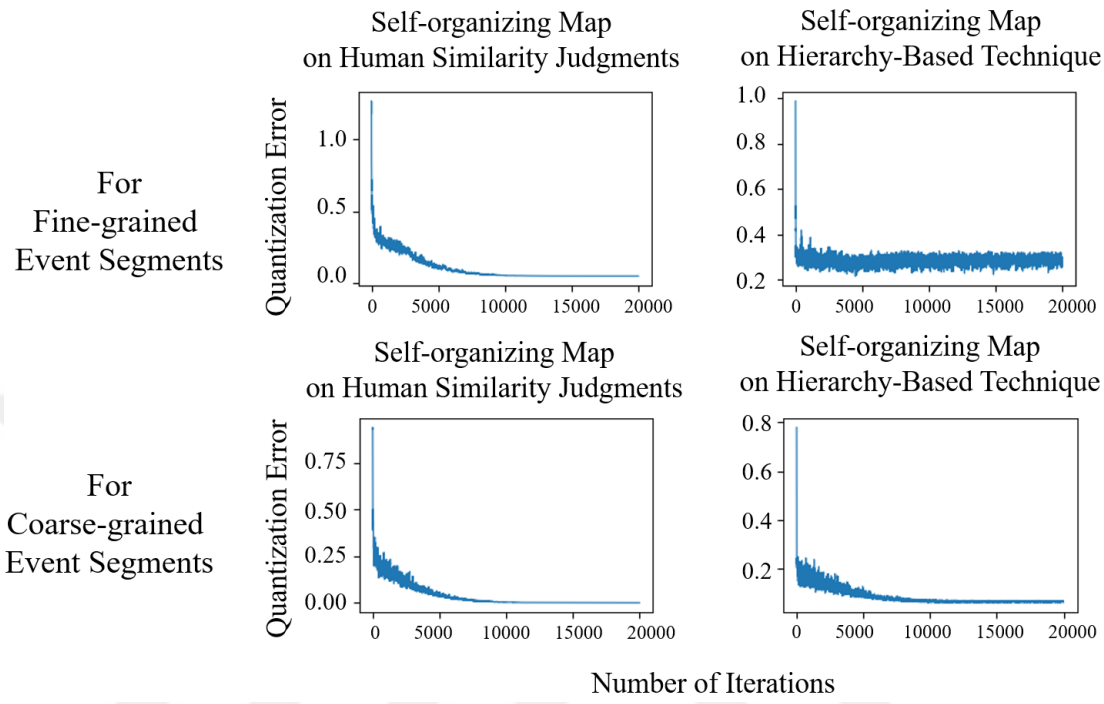
HYPERPARAMETERS OF SELF-ORGANIZING MAPS

Table 3. Hyperparameters of Self-Organizing Maps

Parameters	Fine-grained Events	Coarse-grained Events
Neurons	(4, 5)	(4, 4)
Sigma (Radius)	0.6	0.6
Learning rate	1	1
Topology	Hexagonal	Hexagonal
Neighborhood Function	Gaussian	Gaussian

APPENDIX J

TRAINING OF SELF-ORGANIZING MAPS



Training performances of self-organizing maps:

Quantization errors representing the difference between data and its representation on a self-organizing network. It can be seen from figures that trained self-organizing maps learned to represent data gradually.

APPENDIX K

PSEUDOCODE FOR LOOMING OPERATION

```
FUNCTION GenerateRadiuses(Markers, Frames, interval, maxSize, minSize)
    Initialize radiusesofAllMarkers <- []
    FOR each marker in Markers
        Initialize radiuses <- []
        Append(radiuses, minSize)
        WHILE Length(radiuses) <= Frames:
            Initialize d <- U[0, 1]
            Initialize currentSize <- radiuses[-1] # the last radius appended to the radiuses
            IF d <= 0.45
                radiusesAdded <- Interpolate(currentSize, maxSize, interval) # increase radius
            ELSE IF d <= 0.90
                radiusesAdded <- Interpolate(currentSize, minSize, interval)  # reduce radius
            ELSE
                radiusesAdded <- Interpolate(currentSize, currentSize, interval)  # keep radius constant
            Append(radiuses, radiusesAdded)
        ENDWHILE
        Append(radiusesofAllMarkers, radiuses)
    ENDFOR
    RETURN radiusesofAllMarkers
```

APPENDIX L

PSEUDOCODE FOR MATCHING CHANGES

```
FUNCTION boundaryCheck(X, Y, xLocations, yLocations)
    xmax, xmin <- findMaximum(xLocations), findMinimum(xLocations)
    ymax, ymin <- findMaximum(yLocations), findMinimum(yLocations)

    IF xmax >= x >= xmin
        xBool <- True
    ELSE
        xBool <- False
    ENDIF

    IF ymax >= y >= ymin
        yBool <- True
    ELSE
        yBool <- False
    ENDIF

    RETURN xBool, yBool
```

```
FUNCTION returningMaximumChange(oldLocationx, oldLocationy, requiredChange,
                                xLocations, yLocations, possibleLocations=[])
    directions <- [1, -1]
    dimensions <- [x, y]
    pXlocations, pYlocations <- [], []

    FOR each dimension in dimensions do
        FOR each direction in directions do
            Append(pXlocations, oldLocationx + direction*requiredChange/2)
            Append(pYlocations, oldLocationy + direction*requiredChange/2)
        ENDFOR
    ENDFOR

    Initialize maxLocations <- []
    FOR x in pXlocations do
        FOR y in pYlocations
            xBool, yBool = boundaryCheck(X, Y, xLocations, yLocations)
            IF xBool AND yBool
                Append(maxLocations, [xBool, yBool])
            ENDIF
        ENDFOR
    ENDFOR

    IF requiredChange < 0.0000001
        RETURN [[oldLocationx, oldLocationy]]
    ELSE IF LENGTH(maxLocations) == 0
        requiredChangeModulated <- requiredChange*0.9
        RETURN returningMaximumChange(oldLocationx, oldLocationy,
                                      requiredChangeModulated, xLocations,
                                      yLocations, possibleLocations)
    ELSE
        RETURN maxLocations
```

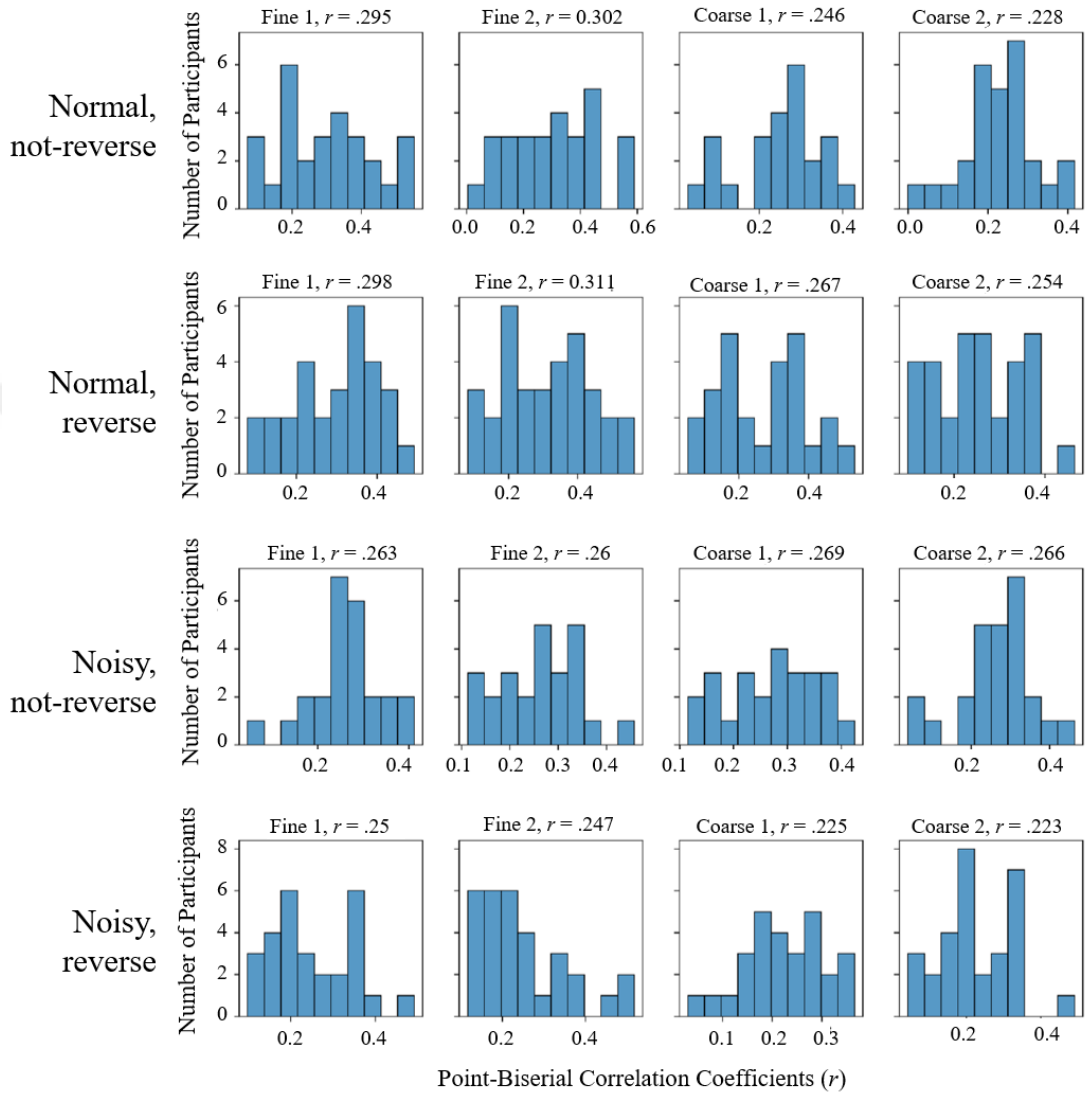
```
FUNCTION generateTrajectory(originX, originY, requiredChanges, xLocations, yLocations)
    xLocationsMarker, yLocationsMarker <- [originX], [originY]
    FOR x, y, required_change in xLocations, yLocations, requiredChanges do
        oldLocationx, oldLocationy = xLocationsMarker[-1], yLocationsMarker[-1]
        maxLocations = returningMaximumChange(oldLocationx, oldLocationy,
                                              required_change, xLocations, yLocations)
        Sample newLocationX, newLocationY from maxLocations
            Append(xLocationsMarker, newLocationX)
            Append(yLocationsMarker, newLocationY)
    ENDFOR
    RETURN xLocationsMarker, yLocationsMarker


FUNCTION computeDifference(normalCoordinates, noisyCoordinates)
    normalDifference <- Sum(Absolute(Difference(normalCoordinates)))
    noisyDifference <- Sum(Absolute(Difference(noisyCoordinates)))
    requiredChanges <- noisyDifference - normalDifference
    RETURN requiredChanges


FUNCTION matchChanges(normalCoordinates, noisyCoordinates, markerNames)
    FOR marker in MarkerNames do
        requiredChanges <- computeDifference(normalCoordinates, noisyCoordinates)
        xLocations, yLocations <- noisyCoordinates
        originX, originY <- xLocations[0], ylocations[0]
        xLocationsMarker, yLocationsMarker <- generateTrajectory(originX, originY,
                                              requiredChanges,
                                              xLocations, yLocations)
        noisyCoordinatesMarker <- xLocationsMarker, yLocationsMarker
    ENDFOR
    RETURN normalCoordinates, noisyCoordinates
```

APPENDIX M

INTRAGROUP CORRELATIONS IN PSYCHOLOGICAL EXPERIMENTS



APPENDIX N

DATA EXCLUSIONS

Sensory Reliability	Expectation	Event Granularity	Collected Data	Experiment Stuck Check	Attention Question	Intragroup Correlations	Outliers	Remaining Data
Noisy	Not-inverted	Coarse 1	34	7	3	0	1	23
Noisy	Not-inverted	Coarse 2	34	7	2	2	1	23
Noisy	Not-inverted	Fine 1	34	7	3	1	1	23
Noisy	Not-inverted	Fine 2	34	7	2	0	1	24
Noisy	Inverted	Coarse 1	36	5	4	2	0	25
Noisy	Inverted	Coarse 2	36	5	2	3	0	26
Noisy	Inverted	Fine 1	36	5	4	1	1	25
Noisy	Inverted	Fine 2	36	5	0	0	1	30
Normal	Not-inverted	Coarse 1	37	6	5	3	1	22
Normal	Not-inverted	Coarse 2	37	6	1	3	1	26
Normal	Not-inverted	Fine 1	37	6	1	1	0	29
Normal	Not-inverted	Fine 2	37	6	1	4	0	26
Normal	Inverted	Coarse 1	39	4	9	2	1	24
Normal	Inverted	Coarse 2	39	4	1	2	1	31
Normal	Inverted	Fine 1	39	4	6	1	1	27
Normal	Inverted	Fine 2	39	4	1	3	0	31
			584	88	45	28	11	415

Data exclusions:

Data removal because of techniques employed for maintaining data reliability

REFERENCES

Alaerts, K., Nackaerts, E., Meyns, P., Swinnen, S. P., & Wenderoth, N. (2011). Action and emotion recognition from point light displays: an investigation of gender differences. *PLoS ONE*, *6*(6), 20989. doi:10.1371/journal.pone.0020989

Andrychowicz, M., Wolski, F., Ray, A., Schneider, J., Fong, R., Welinder, P., … Zaremba, W. (2018). Hindsight experience replay. *30th International Conference on Advances in Neural Information Processing Systems, 30(1)*, 1-11

Baldwin, D. A., Baird, J. A., Saylor, M. M., & Clark, M. A. (2001). Infants parse dynamic action. *Child Development*, *72*(3), 708–717. doi:10.1111/1467-8624.00310

Basgol, H., Ayhan, I., & Ugur, E. (2021). Time perception: a review on psychological, computational and robotic models. *IEEE Transactions on Cognitive and Developmental Systems*, *1*(1), 1. doi:10.1109/TCDS.2021.3059045

Bastos M., A., Usrey Martin, W., Adams A., R., Mangun R., G., Fries, P., & Friston J., K. (2012). Canonical microcircuits for predictive coding. *Neuron*, *76*(4), 695–711. doi:10.1016/j.neuron.2012.10.038

Bengio, Y., Courville, A., & Vincent, P. (2014). Representation learning: a review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(8), 1798–1828. doi:10.1109/TPAMI.2013.50

Bertenthal, B. I., & Pinto, J. (1994). Global processing of biological motions. *Psychological Science*, *5*(4), 221–225. doi:10.1111/j.1467-9280.1994.tb00504.x

Blom, T., Feuerriegel, D., Johnson, P., Bode, S., & Hogendoorn, H. (2020). Predictions drive neural representations of visual events ahead of incoming sensory information. *Proceedings of the National Academy of Sciences*, *117*(13), 7510–7515. doi:10.1073/pnas.1917777117

Blough, D. S. (2001). The perception of similarity. In R. G. Cook (Ed.), *Avian visual cognition* (pp. 23–25). Comparative Cognition Press. Retrieved from www.pigeon.psy.tufts.edu/avc/dblough/

Butz, M. V, Bilkey, D., Humaidan, D., Knott, A., & Otte, S. (2019). Learning, planning, and control in a monolithic neural event inference architecture. *Neural Networks*, *117*(1), 135–144. doi:10.1016/j.neunet.2019.05.001

Chalk, M., Seitz, A. R., & Series, P. (2010). Rapidly learned stimulus expectations alter perception of motion. *Journal of Vision*, *10*(8), 2. doi:10.1167/10.8.2

Chollet, F. (2015). Keras. Retrieved from https://keras.io

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences, 36*(3), 181–204. doi:10.1017/s0140525x12000477

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences* (2nd ed.). New York, USA: Routledge. doi:10.4324/9780203771587

Cousineau, D. (2005). Confidence intervals in within-subject designs: A simpler solution to Loftus and Masson's method. *Tutorials in Quantitative Methods for Psychology*, *1*(1), 42–45. doi:10.20982/tqmp.01.1.p042

Cutting, J. E., Brunick, K. L., & Candan, A. (2012). Perceiving event dynamics and parsing Hollywood films. *Journal of Experimental Psychology: Human Perception and Performance*, *38*(6), 1476–1490. doi:10.1037/a0027737

Day, S. B., & Bartels, D. M. (2008). Representation over time: The effects of temporal distance on similarity. *Cognition*, *106*(3), 1504–1513. doi:10.1016/j.cognition.2007.05.013

De Lange, F. P., Heilbron, M., & Kok, P. (2018). How do expectations shape perception? *Trends in Cognitive Sciences*, *22*(9), 764–779. doi:10.1016/j.tics.2018.06.002

Deselaers, T., & Ferrari, V. (2011). Visual and semantic similarity in ImageNet. *Conference on Computer Vision and Pattern Recognition (CVPR), 12(2)*, 1777-1784. doi:10.1109/cvpr.2011.5995474

Dias, C., & Dimiccoli, M. (2011). Learning event representations by encoding the temporal context. *European Conference on Computer Vision (ECCV), 1(2)*, 1-10. doi:10.1109/cvpr.2011.5995474

Dimiccoli, M., & Wendt, H. (2020). Learning event representations for temporal segmentation of image sequences by dynamic graph embedding. *IEEE Transactions on Image Processing*, *30(1)*, 1476–1486. doi:10.1109/TIP.2020.3044448

DuBrow, S., & Davachi, L. (2014). Temporal memory is shaped by encoding stability and intervening item reactivation. *Journal of Neuroscience*, *34*(42), 13998–14005. doi:10.1523/JNEUROSCI.2535-14.2014

Eisenberg, M. L., Zacks, J. M., & Flores, S. (2018). Dynamic prediction during perception of everyday events. *Cognitive Research: Principles and Implications*, *3*(1), 1–12. doi:10.1186/s41235-018-0146-z

Eslami, S. M. A., Jimenez Rezende, D., Besse, F., Viola, F., Morcos, A. S., Garnelo, M., … Hassabis, D. (2018). Neural scene representation and rendering. *Science*, *360*(6394), 1204–1210. doi:10.1126/science.aar6170

Feldman, H., & Friston, K. J. (2010). Attention, uncertainty, and free-energy. *Frontiers in Human Neuroscience*, *4*(1), 215. doi:10.3389/fnhum.2010.00215

Fivush, R., Kuebli, J., & Clubb, P. A. (1992). The structure of events and event representations: a developmental analysis. *Child Development*, *63*(1), 188. doi:10.2307/1130912

Fountas, Z., Sylaidi, A., Nikiforou, K., Seth, A. K., Shanahan, M., & Roseboom, W. (2020). A predictive processing model of episodic memory and time perception. *BioRxiv*, *1*(1), 1-23. 2020.02.17.953133. doi:10.1101/2020.02.17.953133

Franklin, N. T., Norman, K. A., Ranganath, C., Zacks, J. M., & Gershman, S. J. (2020). Structured event memory: A neuro-symbolic model of event cognition. *Psychological Review*, *127*(3), 327–361. doi:10.1037/rev0000177

French, R. M. (1999). Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences*, *3*(4), 128–135. doi:10.1016/S1364-6613(99)01294-2

Funke, C. M., Borowski, J., Stosio, K., Brendel, W., Wallis, T. S. A., & Bethge, M. (2021). Five points to check when comparing visual perception in humans and machines. *Journal of Vision*, *21*(3), 1–23. doi:10.1167/jov.21.3.16

Giese, M. A., & Poggio, T. (2003). Neural mechanisms for the recognition of biological movements. *Nature Reviews Neuroscience*, *4*(3), 179–192. doi:10.1038/nrn1057

Giorgino, T. (2009). Computing and visualizing dynamic time warping alignments in R: The dtw package. *Journal of Statistical Software*, *31*(7), 1–24. doi:10.18637/jss.v031.i07

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. Cambridge, MA: MIT Press. Retrieved from www.deeplearningbook.org/

Gosling, S. D., & Mason, W. (2015). Internet research in psychology. *Annual Review of Psychology*, *66*, 877–902. doi:10.1146/annurev-psych-010814-015321

Graziano, W. G., Moore, J. S., & Collins, J. E. (1988). Social cognition as segmentation of the stream of behavior. *Developmental Psychology*, *24*(4), 568–573. doi:10.1037/0012-1649.24.4.568

Gumbsch, C., Butz, M. V., & Martius, G. (2019). Autonomous identification and goal-directed invocation of event-predictive behavioral primitives. *IEEE Transactions on Cognitive and Developmental Systems*, *1*(1), 1. doi:10.1109/TCDS.2019.2925890

Gumbsch, C., Kneissler, J., & Butz, M. V. (2016). Learning behavior-grounded event segmentations. *38th Annual Meeting of the Cognitive Science Society*, 1(1), 1787-1792.

Gumbsch, C., Otte, S., & Butz, M. V. (2017). A computational model for the dynamical learning of event taxonomies. *39th Annual Meeting of the Cognitive Science Society*, 1(1), 452-457.

Hard, B. M., Recchia, G., & Tversky, B. (2011). The shape of action. *Journal of Experimental Psychology: General*, *140*(4), 586–604. doi:10.1037/a0024310

Hard, B. M., & Tversky, B. (2003). Segmenting ambiguous events. *25th Annual Meeting of the Cognitive Science Society*, 1(1), 781-786.

Hard, B. M., Tversky, B., & Lang, D. S. (2006). Making sense of abstract events: Building event schemas. *Memory & Cognition*, *34*(6), 1221–1235. doi:10.3758/bf03193267

Haushofer, J., Livingstone, M. S., & Kanwisher, N. (2008). Multivariate patterns in object-selective cortex dissociate perceptual and physical shape similarity. *PLoS Biology*, *6*(7), 1459–1467. doi:10.1371/journal.pbio.0060187

Hebart, M. N., Zheng, C. Y., Pereira, F., & Baker, C. I. (2020). Revealing the multidimensional mental representations of natural objects underlying human similarity judgements. *Nature Human Behaviour*, *4*(11), 1173–1185. doi:10.1038/s41562-020-00951-3

Heider, F., & Simmel, M. (1944). An experimental study of apparent behavior. *The American Journal of Psychology*, *57*(2), 243–259. doi:10.2307/1416950

Hemeren, P. E., & Thill, S. (2011). Deriving motor primitives through action segmentation. *Frontiers in Psychology*, *1*(1), 243. doi:10.3389/fpsyg.2010.00243

Hohwy, J., Hebblewhite, A., & Drummond, T. (2021). Events, event prediction, and predictive processing. *Topics in Cognitive Science*, *13*(1), 252–255. doi:10.1111/tops.12491

Huff, M., Papenmeier, F., & Zacks, J. M. (2012). Visual target detection is impaired at event boundaries. *Visual Cognition*, *20*(7), 848–864. doi:10.1080/13506285.2012.705359

Hunt, A. R., & Halper, F. (2008). Disorganizing biological motion. *Journal of Vision*, *8*(9), 12. doi:10.1167/8.9.12

Johansson, G. (1973). Visual perception of biological motion and a model for its analysis. *Perception & Psychophysics*, *14*(2), 201–211. doi:10.3758/bf03212378

Jozwik, K. M., Kriegeskorte, N., Storrs, K. R., & Mur, M. (2017). Deep convolutional neural networks outperform feature-based but not categorical models in explaining object similarity judgments. *Frontiers in Psychology*, *8*(1), 1726. doi:10.3389/fpsyg.2017.01726

Kohonen, T. (1990). The self-organizing map. *Proceedings of the IEEE*, *78*(9), 1464–1480. doi:10.1109/5.58325

Kok, P., Brouwer, G. J., Van Gerven, M. A. J., & De Lange, F. P. (2013). Prior expectations bias sensory representations in visual cortex. *Journal of Neuroscience*, *33*(41), 16275–16284. doi:10.1523/jneurosci.0742-13.2013

Kominsky, J. F., Baker, L., Keil, F. C., & Strickland, B. (2021). Causality and continuity close the gaps in event representations. *Memory & Cognition*, *49*(3), 518–531. doi:10.3758/s13421-020-01102-9

Kriegeskorte, N., & Mur, M. (2012). Inverse MDS: Inferring dissimilarity structure from multiple item arrangements. *Frontiers in Psychology, 3*(1), 245. doi:10.3389/fpsyg.2012.00245

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 25(1), 1097–1105. doi:10.1145/3065386

Kurby, C. A., & Zacks, J. M. (2008). Segmentation in the perception and memory of events. *Trends in Cognitive Sciences*, *12*(2), 72–79. doi:10.1016/j.tics.2007.11.004

Markus, H., Smith, J., & Moreland, R. L. (1985). Role of the self-concept in the perception of others. *Journal of Personality and Social Psychology*, *49*(6), 1494–1512. doi:10.1037/0022-3514.49.6.1494

Metcalf, K., & Leake, D. B. (2017). Modelling unsupervised event segmentation: learning event boundaries from prediction errors. *39th Annual Meeting of the Cognitive Science Society*, 1(1), 2717-2722.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *26th International Conference on Neural Information Processing Systems, 2(1),* 3111-3119

Neisser, U. (1991). *Cognition and reality: principles and implications of cognitive psychology*. New York: W. H. Freeman.

Nery, B., & Ventura, R. (2011). A dynamical systems approach to online event segmentation in cognitive robotics. *Journal of Behavioral Robotics*, *2*(1), 18–24. doi:10.2478/s13230-011-0011-y

Newtson, D. (1973). Attribution and the unit of perception of ongoing behavior. *Journal of Personality and Social Psychology*, *28*(1), 28–38. doi:10.1037/h0035584

Newtson, D., & Engquist, G. (1976). The perceptual organization of ongoing behavior. *Journal of Experimental Social Psychology, 12*(5), 436–450. doi:10.1016/0022-1031(76)90076-7

Newtson, D., Engquist, G. A., & Bois, J. (1977). The objective basis of behavior units. *Journal of Personality and Social Psychology*, *35*(12), 847–862. doi:10.1037/0022-3514.35.12.847

Ney, H. J., & Ortmanns, S. (1999). Dynamic programming search for continuous speech recognition. *IEEE Signal Processing Magazine*, *16*(5), 64–83. doi:10.1109/79.790984

Noble, K., Glowinski, D., Murphy, H., Jola, C., Mcaleer, P., Darshane, N., … Pollick, F. E. (2014). Event segmentation and biological motion perception in watching dance. *Art & Perception*, *2*(2), 59–74. doi:10.1163/22134913-00002011

Nosofsky, R. M. (1992). Similarity scaling and cognitive process models. *Annual Review of Psychology*, *43*(1), 25–53. doi:10.1146/annurev.ps.43.020192.000325

O'Reilly, J. X. (2013). Making predictions in a changing world-inference, uncertainty, and learning. *Frontiers in Neuroscience*, *7*(7), 105. doi:10.3389/fnins.2013.00105

Ólafsdóttir, H. F., Bush, D., & Barry, C. (2018). The role of hippocampal replay in memory and planning. *Current Biology*, *28*(1), 37–50. doi:10.1016/j.cub.2017.10.073

Pavlova, M., & Sokolov, A. (2003). Prior knowledge about display inversion in biological motion perception. *Perception*, *32*(8), 937–946. doi:10.1068/p3428

Peirce, J., Gray, J. R., Simpson, S., Macaskill, M., Höchenberger, R., Sogo, H., … Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, *51*(1), 195–203. doi:10.3758/s13428-018-01193-y

Peterson, J. C., Abbott, J. T., & Griffiths, T. L. (2018). Evaluating (and improving) the correspondence between deep neural networks and human representations. *Cognitive Science*, *42*(8), 2648–2669. doi:10.1111/cogs.12670

Pitt, D. (2020). Mental Representation. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. Retrieved from: plato.stanford.edu/entries/mental-representation/

Plappert, M., Mandery, C., & Asfour, T. (2016). The KIT motion-language dataset. *Big Data*, *4*(4), 236–252. doi:10.1089/big.2016.0028

Radvansky, G. A., & Copeland, D. E. (2006). Walking through doorways causes forgetting: Situation models and experienced space. *Memory & Cognition*, *34*(5), 1150–1156. doi:10.3758/BF03193261

Reynolds, J. R., Zacks, J. M., & Braver, T. S. (2007). A computational model of event segmentation from perceptual prediction. *Cognitive Science*, *31*(4), 613–643. doi:10.1080/15326900701399913

Richmond, L. L., & Zacks, J. M. (2017). Constructing experience: event models from perception to action. *Trends in Cognitive Sciences*, *21*(12), 962–980. doi:10.1016/j.tics.2017.08.005

Rogers, T. T., & McClelland, J. L. (2005). A parallel distributed processing approach to semantic cognition: Applications to conceptual development. In L. Gershkoff-Stowe & D. H. Rakison (Eds.) *Carnegie Mellon Symposia on cognition. Building object categories in developmental time* (pp. 335-387). New Jersey: Lawrence Erlbaum Associates Publishers.

Rosenthal, R. (2011). Meta-analytic procedures for social research. In *Meta-analytic procedures for social research*. New York City: SAGE Publications, Inc. doi:10.4135/9781412984997

Rothfuss, J., Ferreira, F., Aksoy, E. E., Zhou, Y., & Asfour, T. (2018). Deep episodic memory: encoding, recalling, and predicting episodic experiences for robot action execution. *IEEE Robotics and Automation Letters*, *3*(4), 4007–4014. doi:10.1109/LRA.2018.2860057

Schapiro, A. C., Rogers, T. T., Cordova, N. I., Turk-Browne, N. B., & Botvinick, M. M. (2013). Neural representations of events arise from temporal community structure. *Nature Neuroscience*, *16*(4), 486–492. doi:10.1038/nn.3331

Schönemann, P. H. (1966). A generalized solution of the orthogonal procrustes problem. *Psychometrika*, *31*(1), 1–10. doi:10.1007/BF02289451

Schütz-Bosbach, S., & Prinz, W. (2007). Prospective coding in event representation. *Cognitive Processing*, *8*(2), 93–102. doi:10.1007/s10339-007-0167-x

Schwan, S., & Garsoffky, B. (2004). The cognitive representation of filmic event summaries. *Applied Cognitive Psychology*, *18*(1), 37–55. doi:10.1002/acp.940

Sheldon, S., & El-Asmar, N. (2018). The cognitive tools that support mentally constructing event and scene representations. *Memory*, *26*(6), 858–868.

Shen, J., Fu, C., Deng, X., & Ino, F. (2020). A study on training story generation models based on event representations. *The 3rd International Conference on Artificial Intelligence and Big Data*, *3*(1), 210–214. doi:10.1109/ICAIBD49809.2020.9137439

Shepard, R. N. (1980). Multidimensional scaling, tree-fitting, and clustering. *Science*, *210*(4468), 390–398. doi:10.1126/science.210.4468.390

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*(4820), 1317–1323. doi:10.1126/science.3629243

Shepard, R. N., & Arabie, P. (1979). Additive clustering: Representation of similarities as combinations of discrete overlapping properties. *Psychological Review*, *86*(2), 87. doi:10.1037/0033-295X.86.2.87

Speer, N. K., Zacks, J. M., & Reynolds, J. R. (2007). Human brain activity time-locked to narrative event boundaries. *Psychological Science*, *18*(5), 449–455. doi:10.1111/j.1467-9280.2007.01920.x

Stawarczyk, D., Bezdek, M. A., & Zacks, J. M. (2021). Event representations and predictive processing: the role of the midline default network core. *Topics in Cognitive Science*, *13*(1), 164–186. doi:10.1111/tops.12450

Tarhan, L., de Freitas, J., Alvarez, G., & Konkle, T. (2020). Semantic embeddings of verbal descriptions predict action similarity judgments. *Journal of Vision*, *20*(11), 1241.

Tarhan, L., & Konkle, T. (2018). Predicting the behavioral similarity structure of visual actions. *Journal of Vision*, *18*(10), 428.

Troje, N. F., & Basbaum, A. (2008). Biological motion perception. In B. Fritzsch (Ed.), *The senses: A comprehensive reference* (2nd ed., pp. 231–238). Cambridge: Academic Press.

Troje, N. F., & Westhoff, C. (2006). The inversion effect in biological motion perception: evidence for a "life detector"? *Current Biology*, *16*(8), 821–824. doi:10.1016/j.cub.2006.03.022

Tversky, A. (1977). Features of similarity. *Psychological Review*, *84*(4), 327–352. doi:10.1037/0033-295X.84.4.327

Urban, C. J., & Gates, K. (2019). Deep learning: a primer for psychologists. *PsyArXiv*. doi:https://doi.org/10.31234/osf.io/4q8na

van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, *9*(86), 2579–2605.

Wang, J., Cherkassky, V. L., & Just, M. A. (2017). Predicting the brain activation pattern associated with the propositional content of a sentence: Modeling neural representations of events and states. *Human Brain Mapping*, *38*(10), 4865–4881. doi:10.1002/hbm.23692

Wei, P., Zhao, Y., Zheng, N., & Zhu, S.-C. (2016). Modeling 4d human-object interactions for joint event segmentation, recognition, and object localization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *39*(6), 1165–1179.

Wiese, W., & Metzinger, T. (2017). Vanilla PP for Philosophers: A Primer on Predictive Processing. In W. Metzinger, Thomas K. and Wiese (Ed.), *Philosophy and predictive processing* (pp. 1–18). Frankfurt: MIND Group. doi:10.15502/9783958573024

Wilder, D. A. (1978a). Effect of predictability on units of perception and attribution. *Personality and Social Psychology Bulletin*, *4*(2), 281–284. doi:10.1177/014616727800400222

Wilder, D. A. (1978b). Predictability of behaviors, goals, and unit of perception. *Personality and Social Psychology Bulletin*, *4*(4), 604–607.

Wold, S., Esbensen, K., & Geladi, P. (1987). Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, *2*(1–3), 37–52. doi:10.1016/0169-7439(87)80084-9

Zacks, J. M. (2004). Using movement and intentions to understand simple events. *Cognitive Science*, *28*(6), 979–1008. doi:10.1207/s15516709cog2806_5

Zacks, J. M. (2020). Event perception and memory. *Annual Review of Psychology*, *71*(1), 165–191. doi:10.1146/annurev-psych-010419-051101

Zacks, J. M., Braver, T. S., Sheridan, M. A., Donaldson, D. I., Snyder, A. Z., Ollinger, J. M., … Raichle, M. E. (2001). Human brain activity time-locked to perceptual event boundaries. *Nature Neuroscience*, *4*(1), 651–655. doi:https://doi.org/10.1038/88486

Zacks, J. M., Kurby, C. A., Eisenberg, M. L., & Haroutunian, N. (2011). Prediction error associated with the perceptual segmentation of naturalistic events. *Journal of Cognitive Neuroscience*, *23*(12), 4057–4066. doi:10.1162/jocn_a_00078

Zacks, J. M., Speer, N. K., Swallow, K. M., Braver, T. S., & Reynolds, J. R. (2007). Event perception: A mind-brain perspective. *Psychological Bulletin*, 133(1), 273–293. doi:10.1037/0033-2909.133.2.273

Zacks, J. M., Speer, N. K., Swallow, K. M., & Maley, C. J. (2010). The brain's cutting-room floor: Segmentation of narrative cinema. *Frontiers in Human Neuroscience*, *4*(1), 168. doi:10.3389/fnhum.2010.00168

Zacks, J. M., & Swallow, K. M. (2007). Event segmentation. *Current Directions in Psychological Science*, *16*(2), 80–84. doi:10.1111/j.1467-8721.2007.00480.x