A Self-Supervised and Predictive Processing-Based Model of Event Segmentation and Learning

Anonymous CogSci submission

Abstract

Event is a fuzzy term that refers to bounded spatio-temporal units. Events guide behavior to allow adaptation to complex environments. The study of event segmentation investigates mechanisms behind the ability to segment the continuous information flow into discrete units. Event Segmentation Theory states that people predict observed ongoing activities and monitor their prediction errors for event segmentation. In this study, inspired from the principles of Event Segmentation Theory and predictive processing, we introduced a computational model of event segmentation and learning. In order to verify that our method can segment ongoing activity into meaningful parts and learn them via passive observation, we compared the performance of our method with humans for fine and coarse segmentation tasks in two psychological experiments. The results demonstrated that our model not only learned segmented behavioral units accurately but also displayed similar segmentation decisions with human subjects.

Keywords: event segmentation, event learning, point-light displays, action segmentation, predictive processing

Introduction

Humans are subject to continuous flow of information, and have to utilize it for robust, adaptive, and intelligent behavior. In order to utilize this information, they discretize it into meaningful units. Whereas in the spatial realm, this segmentation transforms scenes into meaningful objects; in the temporal realm, the transformation takes place from a sequence of scenes into event units (Zacks, 2020).

Similarity across event segmentation decisions of humans was firstly noted by (Newtson, 1973), who asked participants to segment a movie by pressing a button to detect event boundaries in a procedure called unitization. The results of the study showed that event boundaries had substantial agreement and are stable across time. Furthermore, participants were able to segment activities into small (fine-grained) or large (coarse-grained) events with corresponding task instructions. Event Segmentation Theory (EST) aims to explain mechanisms behind event segmentation (Zacks, 2020). According to the EST, event boundaries are determined by the transiently increasing prediction error, which in turn triggers another event model for the prediction of the next sensory input. From the perspective of predictive processing framework, which asserts that the brain generates models of the environment to predict current sensory input by learning prediction error signals (Wiese & Metzinger, 2017), event models are similar to mental models aiming to predict the current sensory input and reducing the prediction error signals.

There are several computational models of event segmentation developed in cognitive science (Reynolds, Zacks, & Braver, 2007; Gumbsch, Kneissler, & Butz, 2016; Gumbsch, Otte, & Butz, 2017; Franklin, Norman, Ranganath, Zacks, & Gershman, 2019). For example, Reynolds et al. (2007) proposed different sequence models for segmenting human behaviors. However, their proposed models are not able to learn and segment human behaviors in varying hierarchies. Besides, transitions between human behaviors used in modeling purposes are unnatural. Additionally, Gumbsch et al. (2016, 2017) developed models aiming to chunk sensory-motor interaction flow for robotic behavior learning and planning. Representing events by a set of linear models, their models cannot learn and segment non-linear relationships by only one event model. Additionally, assuming the role of sensory-motor interaction for learning and segmenting events, their models do not aim to segment a sequence of activity by passive observation, which, however, is the usual way of investigating human event segmentation (Newtson, 1973; Zacks, 2020). To best of our knowledge, there is no event segmentation model whose segmentation decisions were directly compared with those of human in psychological experiments.

Inspired by the models developed by Gumbsch et al. (2016, 2017), we propose a computational method of event segmentation and learning. As our contribution to the literature, our method is able to produce segments in varying hierarchies via passive observation and learn those segments with the help of multilayer perceptrons, called event models. For segmenting the continuous information flow, our method utilizes prediction error signals and aims to reduce them. By doing so, it conforms to the principles of the EST and predictive processing framework.

In order to test whether segmentation decisions of the proposed method are meaningful, we conducted psychological experiments that are based on unitization paradigm (Newtson, 1973). Psychological experiments involved human behaviors involving complex motion trajectories expressed by point-light displays (PLDs) (Johansson, 1973), which represent a non-pictorial depiction of biological movements. Considering that change and detecting event boundaries are correlated with each other (Hard, Recchia, & Tversky, 2011), we produced two types of videos, namely normal and noisy (i.e, a smoothed version of normal human behaviors). In psychological experiments, we asked participants to segment these

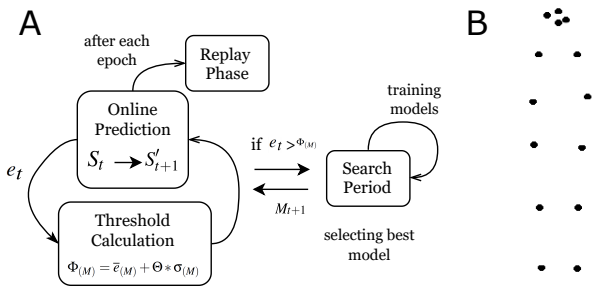

Figure 1: (A) The overview of the proposed method. In the online prediction phase, the current event model makes prediction and the current prediction error is calculated and surprise threshold is computed. If the current prediction error is more than surprise threshold, the system enters to the search period where the best model is returned for online prediction. (B) shows a point-light display representing a human figure.

videos in different hierarchies (i.e., fine and coarse-grained segmentation) and examined whether the noisy video was segmented into the less number of event boundaries than the normal video. We run computational experiments on the same conditions and assessed how well our method performed in event segmentation and learning by the ground-truth data from participants. We were also interested in observing whether our method showed similar biases with human participants against the reduction of the rate of change.

Results of computational experiments showed that, despite occasional mismatches, our method performed well in segmenting human behaviors in varying granularities and learning them. Moreover, ground-truth received from psychological experiments demonstrated that the segmentation decisions of our method are largely consistent with those of humans.

Architecture

Before explaining our proposal, we will explain the computational model by which we are inspired. Gumbsch et al. (2016, 2017) developed a computational model that can autonomously divide a sequence of information and represent events by event models. Event models are a set of linear models. Each of them is responsible from continuously predicting the next sensory input given the current sensory input and corresponding motor command. In their models (Gumbsch et al., 2016, 2017), each forward model generates a prediction for a type of sensory input and calculates a threshold value (so-called surprise threshold) by observed prediction errors. If the prediction error of the current forward model is higher than its surprise threshold (i.e., the current forward model is ineffective for prediction), the system enters into a search period where there is a selection process which takes place between different forward models. Changing the degree of surprise threshold by a value named confidence threshold, they segment events in varying granularities (Gumbsch et al., 2017).

Proposed method

The general overview of our proposal is given in Figure 1A. Rather than exploiting multiple forward models responsible for one type of sensory information (Gumbsch et al., 2016, 2017), considering the complexity of events (i.e, in this context, human behaviors), we decided to use multilayer perceptrons capable of approximating to nonlinear functions. In detail, our proposed architecture has one active model at time t, which is named as event model. M_t is responsible from predicting the change observed in the sensory input. The predicted sensory observation is given by

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

Given the observations, M_t continuously makes predictions, learns from prediction error by changing neural network weights and stores prediction errors.

Search period: event model training and switching Similar to Gumbsch et al. (2016, 2017), surprise threshold Φ is calculated by the rolling mean of stored prediction errors $\overline{e}_{(M)}$ and of the variance $\sigma_{(M)}$. The event threshold Θ regulates the coarseness of the event to be segmented. $Phi_{(M)}$ is calculated by

$$\Phi_{(M)} = \overline{e}_{(M)} + \Theta * \sigma_{(M)}$$

If the prediction error of an event model is greater than $\Phi_{(M)}$, the algorithm enters into the search period, which means that the current event model might not be suitable for predicting the next sensory observation. At the start of the search period, a potential new event model is generated, and all available event models are trained for rehearsal duration corresponding to the number of epochs used in the search period. The training set is sampled from the list formed by $S_{t:t+n}$ and M_{i_S}, representing timesteps that have been predicted by M_i to avoid from the catastrophic forgetting. In short, operations done in the search period finds/generates the most suitable model for corresponding timesteps.

Memory range and replay At the end of each epoch, event models are removed if they were not used for m epochs (i.e, memory range). A replay phase is added at the end of each epoch to avoid catastrophic forgetting, reduce training time, and foster memory consolidation. In the replay phase, all event models, which are used in the recent epoch, are trained by M_{i_S} and Φ is updated. Replay phase is observed in hippocampal regions for memory consolidation (Ólafsdóttir, Bush, & Barry, 2018), and suggested as a technique to stabilize the training process of reinforcement learning agents (Andrychowicz et al., 2017).

Method

Dataset

To assess the capability of the proposed method in segmenting and learning events, we prepared a sequence of human actions depicted by PLDs.

Point-light displays (PLDs) Animals have a strong tendency towards biological movements (Troje, 2008; Johansson, 1973) and humans can perceive biological motion even from moving dots (Johansson, 1973). With the help of relative movement of points, humans perceive various kinds of human movements, emotions, actions, and gender (Troje, 2008). An example point-light display is given in Figure 1B. Representing data as PLDs brings efficiency by reducing the dimensionality of data and processing time.

Human behaviors We took behaviors used in our experiments from the KIT Motion-Language Dataset (Plappert, Mandery, & Asfour, 2016). The activity includes 12 human behaviors, such as walking, jumping, picking an object, sitting on a chair, searching for an object. X and Y dimensions of 14 markers were used to represent behaviors as PLDs, and min-max normalization operation was applied to each behavior taken from the dataset. Behaviors were added back to back through interpolating the marker positions in the point-light display. Thus, a 24 Hz and 270-second video of complex human behavior represented by PLDs was created.

Psychological experiments

In this study, we used the unitization paradigm and determined two experimental conditions, event granularity and sensory reliability. For fine-grained (smaller) events, we asked participants to detect the shortest, natural and meaningful events, whereas, for the latter, they were asked for the longest (larger) ones.

There is an intrinsic relationship between the change and event segmentation. Hard et al. (2011) found that changes at event boundaries are more numerous than those at other frames. Thus, a reduction in change should also decrease the number of event boundaries. From the perspective of predictive coding, the relative reliability of expectations and sensory input determines perception. Ambiguous sensory input (i.e., noisy video) reduce the effect of prediction error (de Lange, Heilbron, & Kok, 2018), from the perspective of the EST, and therefore, decrease the number of perceived event boundaries. In order to verify this, besides event granularity, sensory reliability was determined to be an experimental condition, having two levels as normal and noisy videos. For creating noisy videos, we lowered the degree of change between frames by applying Gaussian white noise, which results in very smooth behaviors.

For both experimental conditions, namely sensory reliability (normal input, noisy input) and event granularity (fine-grained, coarse-grained segmentation), we recruited 19 participants (9 female) for a within-subject design. Each participant firstly attended to normal and then noisy level, where the order of granularity was randomized. For each granularity level, we showed participants the movie twice, referred to as the first and the second observations. In the second observation, participants were asked to segment events in the shortest or longest possible way in accord with the granularity level. It is known that observing an activity more than once results in coarser and finer segmentations (Hard et al., 2011). Throughout the paper, we referred to two observations by numbers. For example, the second observation of fine-grained segmentation is called Fine 2. The experiment was prepared in Psychopy3 (Peirce et al., 2019) and conducted on an online platform named Pavlovia.

Table 1: Hyperparameters of the computational model

Parameters	Fine	Coarse
Event threshold	1.5	4.0
Error window	5	50
Number of timesteps	5	15
Rehearsal	200	300
Replay	2000	2500
Number of epochs	10	10
Memory range	1	1
Activation functions	ReLu	ReLu
Optimizer	Adam	Adam
Learning rate	0.0001	0.0001
Batch size	12	12
Hidden layers	(256, 256, 128, 64)	(1024, 512, 128)

Computational experiments

We ran the same method four times (sensory reliability x segmentation granularity) by changing the hyperparameters for segmentation granularity condition and used the same hyperparameters for sensory reliability to observe the effect of reduced change on segmentation decisions. Therefore, we aimed at testing (1) whether our model captures human event segmentation decisions and (2) is affected by the reduction of the rate of change similarly.

The hyperparameters are given in Table 1. Since the proposed method requires deciding between all available models, the computational requirements rise exponentially as the number of events increases. For this reason, we decided to apply the method to each coarse-grained segment separately, which reduced the number of searching periods and events to be considered.

Results

Results of psychological experiments

We checked whether the number of segmentation decisions of participants were in accord with the granularity condition. Whereas data of three participants were excluded because of producing less segments in the fine-grained level, data of one other participant was excluded as the outlier of Coarse 1 (the first observation of coarse-grained level) ($z > 2.58$). The number of segments produced by participants is given in Figure 2. The results from the fine (M = 63.3, SE = 8.24) and coarse (M = 14.06, SE = 1.19) segmentation levels indicated that the number of segments in fine- was more than in coarse-grained segmentation, $t(29) = 6.25$, $p < .001$, as expected.

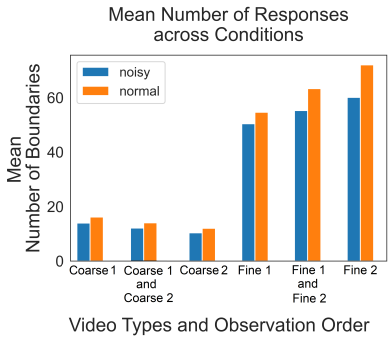

Figure 2: The mean number of boundaries detected in different experimental conditions.

Participants whose data were excluded for normal level were not considered for further analyses. The results from the fine ($M = 55.23$, $SE = 6.79$) and coarse ($M = 12.13$, $SE = 1.48$) segmentation levels indicated that the number of boundaries was smaller in fine- than in coarse-grained segmentation condition, $t(29) = 6.91$, $p < .001$, as expected. As for the relationship between normal and noisy levels, the two-way ANOVA analysis revealed that the main effect of sensory reliability on the number of segmentation decisions was insignificant ($F(1, 112) = 0.84$, $p = .2$.

In order to make a comparison between psychological and computational experiments, we need to find out group agreements over event boundaries, which usually are found by detecting bins receiving responses more than one standard deviation above the mean (Newtson, 1973). We observed that group agreements were sensitive to different bins (1-sec or 2-sec bins) and data sources (whether Coarse 1 or Coarse 2, or both of them are used together). To find the best parameters for comparison, we need to compute a value representing agreement of participants about their group-based segmentation decisions. For computing this kind of value, by varying bin sizes and data sources, we firstly detected group-based segmentation decisions. Secondly, for each group decision, we computed the agreement of participants to that group decision by Cohen's Kappa score, resulting in different Cohen's

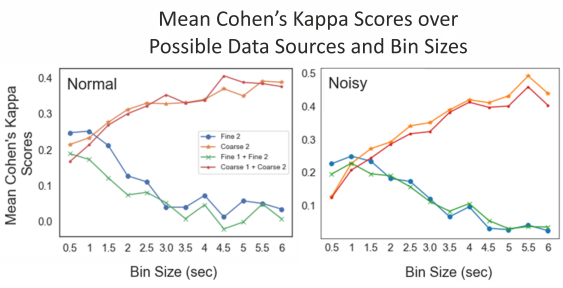

Figure 3: Group agreements over possible data sources and bin sizes.

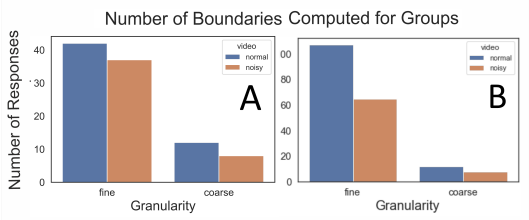

Figure 4: (A) shows the number of boundaries computed when Cohen's Kappa scores are maximized and (B) shows the number of boundaries when fine-grained responses of the noisy video are separated into 0.5-sec bins.

Kappa scores for each participant for each group decision. Thirdly, we computed mean Cohen's Kappa scores of participants to evaluate the group decision. To find the best parameters, we selected the bin size and data sources by maximizing this value. The parameters maximizing mean Cohen's Kappa scores for normal and noisy videos are shown in Figure 3.

For the normal level, we found that the best mean agreement scores were received in Fine 2 with 1-sec bins and in Coarse 1 and Coarse 2 with 4.5-sec bins. For the noisy level, scores were maximized by Fine 2 with 1-sec bins and Coarse 2 with 5.5-sec bins. The number of boundaries detected by groups is given in Figure 4A when their mean Cohen's Kappa scores are maximized. Considering the slight change in Kappa score and the importance of temporal resolution for detecting the true fine boundaries, we decided to use 0.5-sec bins for fine-grained segmentation decisions of both levels (Figure 4B). For the accepted parameters, for normal videos, the group detected 107 fine segments by marking bins receiving responses (*cut-off* = 3.77, $M = 2.0$, $SD = 1.76$) more than one standard deviation above the mean and 12 coarse segments considering bins above the cut-off value (*cut-off* = 13.90, $M = 7.03$, $SD = 6.84$). For noisy videos, 65 fine-grained boundaries (*cut-off* = 3.25, $M = 1.66$, $SD = 1.59$) and 8 coarse-grained boundaries (*cut-off* = 6.57, $M = 3.16$, $SD = 3.41$) were determined. Despite the statistical test that does not reveal a significant difference between normal and noisy videos, the difference in the numbers of group decisions on fine-grained breakpoints is striking.

Results of computational models

We conducted a computational analysis over normal and noisy videos. Figure 5 shows that our method successfully monitors prediction error signals and used them to discover natural, meaningful and identifiable segments. The number of segments detected for each condition by humans and computational model is given in Figure 6. For the normal video, the computational method detected 24 coarse-grained and 87 fine-grained boundaries. For the noisy video, the method revealed 17 coarse-grained and 95 fine-grained boundaries.

We also showed segmentation decisions of humans and computational models on X and Y trajectories of markers of PLDs (Figure 7A for normal and Figure 7B for noisy videos).

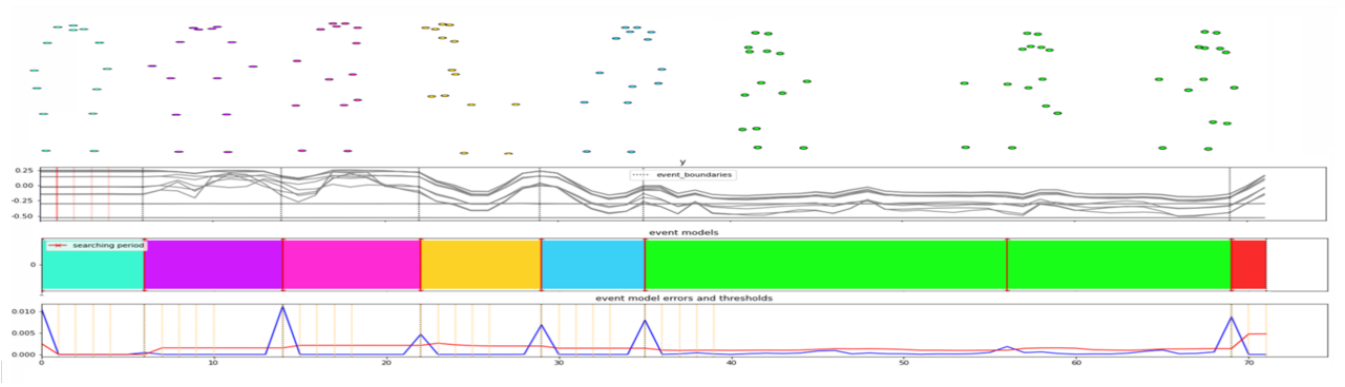

Figure 5: (A) represents the behavior as PLDs format that can be identified as locating objects from one place to another (detected event segments: standing up, bringing hands to wrists, leaning right, leaning left, leaning right, small bending moves, and occasional hand movements to pick up and place the objects), (B) Y coordinates of points. (C) shows event transitions signaled by prediction errors and surprise thresholds (E), marked by blue and red, respectively.

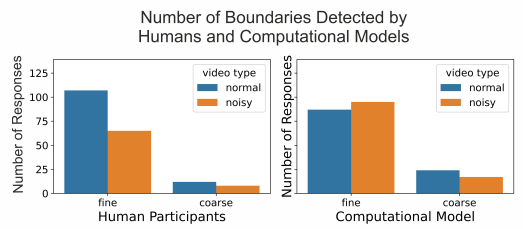

Figure 6: The number of boundaries detected by groups and computational models.

Blue and red lines on Figure 7 represent locations of fine- and coarse-grained boundaries, respectively. The lines at the top of the graph represent segmentation decisions of humans, while the lines on the bottom represent those of computational models.

Despite occasional mismatches between humans and our model, the coarse-grained segmentation decisions are similar to the decisions of humans for the normal video. On the other hand, our method did not capture some fine-grained boundaries detected by humans. For example, between frames 380 and 600 where the walking behavior takes place were not segmented similar to humans. At the same time, it oversegmented certain activities such as the push-upping behavior between frames 580 and 820. Being familiar with the walking activity, participants might have tended to segment it into finer segments compared to push-upping. For the coarse segmentation of noisy videos, our model tended to miss some boundaries and oversegmented certain behaviors. For example, push-upping activity taken place from frames 580 to 820 was segmented into five events. When it comes to fine-grained segmentation of noisy videos, despite the tendency of oversegmentation, our model captured segmentation decisions of participants. Despite slight differences in its responses, our model segmented both videos similarly. For example, it generated occasional fine-grained segments from frames 380 to 580 in the normal video compared to the same range in the noisy video.

Despite the accurate and meaningful segments produced by our model, we did not observe the expected effect of noisy videos on segmentation decisions. The surprise threshold based on the rolling mean of prediction errors might have given rise to this situation. The reduced rate of change might have led to a reduction in prediction error, which reduces the degree of surprise threshold indirectly. Therefore, the relative relationship between prediction errors and the surprise threshold levels remain the same. Further research should investigate the effect of the rate of change on boundary decisions of already trained models.

Discussion

In this study, we proposed a self-supervised, predictive processing based method of event segmentation and learning and tested the model with human data collected from the psychological experiments. We showed that the hierarchical segmentation decisions of our method largely match with the results of psychological experiments. However, as we mentioned, our method does not capture the effect of change. Validating the assumptions of EST, the proposed method showed that observing prediction error is not only a plausible but also a working explanation of event segmentation. Involving event models updated by prediction errors, the proposed method connects event segmentation and predictive processing (Butz, Achimova, Bilkey, & Knott, 2020).

The proposed model has certain shortcomings. Firstly, event models in our proposed method are distinct entities used for prediction, despite the fact that various configurations of human behaviors are grouped in one type of action, such as taking a step or raising a hand. Additionally, our method did not consider temporal regularities between actions (or events). Considering the relationship between prediction errors and time perception (Basgol, Ayhan, & Ugur, 2020), our method can be extended to a time perception model.

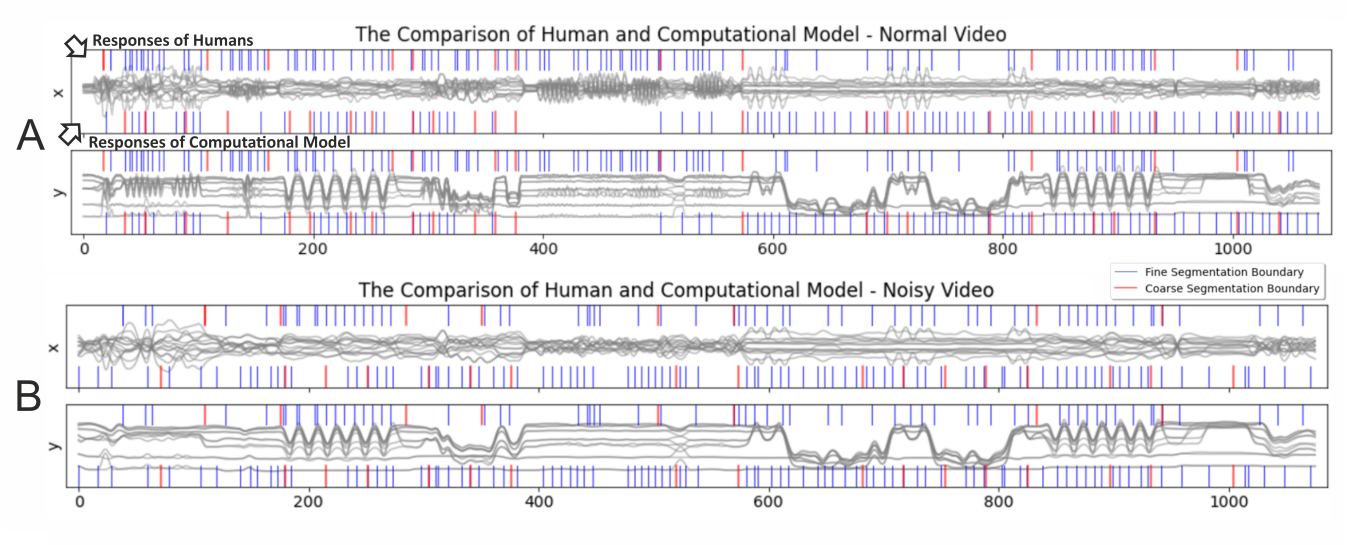

Figure 7: The segmentation decisions of the group and the computational model for normal (A) and noisy (B) videos. The lines at the top of the graph represent group event segmentation decisions, while the lines on the bottom represent the answers of computational models.

Also, representing human behaviors by PLDs brings limitations. For example, PLDs cannot represent human-object or human-human interaction in detail. This problem can be overcome by representing activities by the full-body RGB image. To overcome the simplification of using PLDs, further research can extend the proposed method with the transfer learning approach. Rather than predicting the next sensory input, event models might be trained to predict activations of already trained object identification model like AlexNet (Krizhevsky, Sutskever, & Hinton, 2012).

References

Andrychowicz, M., Wolski, F., Ray, A., Schneider, J., Fong, R., Welinder, P., ... Zaremba, W. (2017). Hindsight experience replay. *arXiv preprint arXiv:1707.01495*.

Basgol, H., Ayhan, I., & Ugur, E. (2020). Time perception: A review on psychological, computational and robotic models. *arXiv e-prints*, arXiv–2007.

Butz, M. V., Achimova, A., Bilkey, D., & Knott, A. (2020). Editors' review and forthcoming topic event-predictive cognition: From sensorimotor via conceptual to language-based structures and processes. *Topics in Cognitive Science*.

de Lange, F. P., Heilbron, M., & Kok, P. (2018). How do expectations shape perception? *Trends in cognitive sciences*, *22*(9), 764–779.

Franklin, N., Norman, K. A., Ranganath, C., Zacks, J. M., & Gershman, S. J. (2019). Structured event memory: a neuro-symbolic model of event cognition. *BioRxiv*, 541607.

Gumbsch, C., Kneissler, J., & Butz, M. V. (2016). Learning behavior-grounded event segmentations. In *Cogsci.*

Gumbsch, C., Otte, S., & Butz, M. V. (2017). A computational model for the dynamical learning of event taxonomies. In *Cogsci.*

Hard, B. M., Recchia, G., & Tversky, B. (2011). The shape of action. *Journal of experimental psychology: General*, *140*(4), 586.

Johansson, G. (1973). Visual perception of biological motion and a model for its analysis. *Perception & psychophysics*, *14*(2), 201–211.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, *25*, 1097–1105.

Newtson, D. (1973). Attribution and the unit of perception of ongoing behavior. *Journal of Personality and Social Psychology*, *28*(1), 28.

Ólafsdóttir, H. F., Bush, D., & Barry, C. (2018). The role of hippocampal replay in memory and planning. *Current Biology*, *28*(1), R37–R50.

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.

Plappert, M., Mandery, C., & Asfour, T. (2016, dec). The KIT motion-language dataset. *Big Data*, *4*(4), 236–252.

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.

Troje, N. F. (2008). Biological motion perception. *The senses: A comprehensive reference*, *2*, 231–238.

Wiese, W., & Metzinger, T. (2017). Vanilla pp for philosophers: A primer on predictive processing. In T. Metzinger & W. Wiese (Eds.), *Philosophy and predictive processing.*

Zacks, J. M. (2020). Event perception and memory. *Annual Review of Psychology*, *71*, 165–191.