supervision for training a computational model, typically a neural network with recurrence for temporal coherence.

We propose a novel computational model [1] based on the concept of perceptual prediction. Defined in cognitive psychology, it refers to the hierarchical process that transforms the current sensory inputs into state representations of the near future that allow for actions. Such representation of the near future enables us to anticipate sensory information based on the current event. This is illustrated in Figure 1. The features were visualized using T-SNE [23] for presentation. The proposed approach has three characteristics. It is hierarchical, recurrent and cyclical. The hierarchical nature of the proposed approach lies in the abstraction of the incoming video frames into features of lower variability that is conducive to prediction. The proposed model is also recurrent. The predicted features are highly dependent on the current and previous states of the network. Finally, the model is highly cyclical. Predictions are compared continuously to observed features and are used to guide future predictions. These characteristics are common working assumptions in many different theories of perception [26], neuro-physiology [11, 7], language processing [34] and event perception[14].

Contributions: The contributions of our proposed approach are three-fold. (1) We are, to the best of our knowledge, the first to tackle the problem of self-supervised, temporal segmentation of videos. (2) We introduce the notion of self-supervised predictive learning for active event segmentation. (3) We show that understanding the spatial-temporal dynamics of events enable the model to learn the visual structure of events for better activity recognition.

2. Related Work

Fully supervised approaches treat event segmentation as a *supervised* learning problem and assign the semantics to the video in terms of labels and try to segment the video into its semantically coherent "*chunks*", with contiguous frames sharing the same label. There have been different approaches to supervised action segmentation such as frame-based labeling using handcrafted features and a support vector machine [19], modeling temporal dynamics using Hidden Markov Models [19], temporal convolutional neural networks (TCN) [27], spatiotemporal convolutional neural networks (CNN) [21] and recurrent networks [29] to name a few. Such approaches often rely on the quantity and quality of the training annotations and constrained by the semantics captured in the training annotations, i.e., a closed world assumption.

Weakly supervised approaches have also been explored to an extent to alleviate the need for large amounts of labeled data. The underlying concept behind weak supervision is to alleviate the need for direct labeling by leveraging accompanying text scripts or instructions as indirect supervision for learning highly discriminant features. There have been two common approaches to weakly supervised learning for temporal segmentation of videos - (1) using script or instructions for weak annotation[6, 10, 3, 24], and (2) following an incomplete temporal localization of actions for learning and inference[16, 29]. While such approaches model the temporal transitions using RNNs, they still rely on enforcing semantics for segmenting actions and hence require some supervision for learning and inference.

Unsupervised learning has not been explored to the same extent as supervised approaches, primarily because label semantics, if available, aid in segmentation. The primary approach is to use clustering as the unsupervised approach using discriminant features[4, 30]. The models incorporate a temporal consistency into the segmentation approach by using either LSTMs [4] or generalized mallows model [30]. Garcia *et al.* [12] explore the use of a generative LSTM network to segment sequences like we do, however, they handle only coarse temporal resolution in life-log images sampled as far apart as 30 seconds. Consecutive images when events change have more variability making for easier discrimination. Besides, they require an iterative training process, which we do not.

3. Perceptual Prediction Framework

In this section, we introduce the proposed framework. We begin with a discussion on the perceptual processing unit, including encoding, prediction and feature reconstruction. We continue with an explanation of the self-supervised approach for training the model, followed by a discussion on boundary detection and adaptive learning. We conclude with implementation details of the proposed approach. It is to be noted that [25] also propose a similar approach based on the Event Segmentation Theory. However, the event boundary detection is achieved using a reinforcement learning paradigm that requires significant amounts of training data and iterations and the approach has only been demonstrated on motion capture data.

3.1. Perceptual Processing

We follow the general principles outlined in the Event Segmentation Theory proposed by Zacks *et al.* [41, 42, 40]. At the core of the approach, illustrated in Figure 2 is a predictive processing platform that encodes a visual input $I(t)$ into a higher level abstraction $I'(t)$ using an encoder network. The abstracted feature is used as a prior to predict the anticipated feature $I'(t+1)$ at time $t+1$. The reconstruction or decoder network creates the anticipated feature, which is used to determine the event boundaries between successive activities in streaming, input video.

[1] Additional results and code can be found at http:\\www.eng.usf.edu\cvprg